FEDERICO RAVENNA
CARL E. WALSH

Screening and Labor Market Flows in a Model with Heterogeneous Workers

We construct a model in which screening of heterogeneous workers by employers plays a central role in determining both the flows into and out of unemployment. Following a negative productivity shock, the share of low-efficiency workers in the pool of unemployed rises, and this composition effect reduces the incentive of firms to post vacancies, lowering job opportunities for *all* workers. Heterogeneity in workers' efficiency amplifies unemployment fluctuations in economies with small gross labor flows and leads to persistent buildups of unemployment and slow recoveries. The composition effect worsens the unemployment–inflation trade-off faced by the monetary authority, leading to very large sacrifice ratios when a fall in productivity primarily affects low-efficiency workers.

JEL codes: E52, E58, J64
Keywords: monetary policy, labor frictions, heterogeneity, unemployment.

WE CONSTRUCT A MONETARY model in which workers are heterogeneous and productivity is worker-specific, resulting in time-varying deviations between the average productivity of employed and unemployed workers. We assume the worker-specific component of productivity, which we label as the worker's *efficiency level*, is unobservable *ex ante* by firms, and job search is nondirected. By interviewing an unemployed worker, firms can observe the worker's efficiency level. Workers with low overall productivity may be interviewed but not hired as firms screen these workers out during the interview process. We show that screening amplifies the volatility of the exit rate from unemployment, capturing the idea that firms become more selective in a recession, reducing the vacancy yield relative to a model with homogeneous levels of efficiency. Following a negative productivity shock, the

The authors would like to thank Wouter den Haan and seminar participants at the 2011 SCG Conference of the Swiss National Bank, the 15th T2M Conference on Theory and Methods in Macroeconomics, the University of Mannheim, McGill University, Paris School of Economics, and the University of Colorado, Boulder, for helpful comments.

FEDERICO RAVENNA *is an Associate Professor, HEC Montreal, Institute of Applied Economics* (*E-mail:* federico.ravenna@hec.ca). CARL E. WALSH *is Professor, University of California, Santa Cruz* (*E-mail:* walshc@ucsc.edu).

Received December 19, 2011; and accepted in revised form April 19, 2012.

Journal of Money, Credit and Banking, Supplement to Vol. 44, No. 2 (December 2012)

share of low-efficiency workers in the pool of unemployed rises, and this composition effect lowers the average productivity of the unemployed relative to the employed. As the share of low-efficiency workers increases, the incentive to post vacancy falls, lowering job opportunities for *all* workers.

The composition effect in our model can act as a powerful mechanism for amplifying the relative volatility of unemployment to output. Separations at the beginning of a recession disproportionately affect low-efficiency workers, and the decline in the job-finding probability is larger for low- relative to high-efficiency workers. Thus the composition effect makes the average total factor productivity (TFP) of unemployed workers more volatile than the TFP of the overall labor force. In fact, measured labor productivity of those who remain employed will *increase* if the recession is driven by a demand shock. We show that a strong amplification effect can obtain even if low-efficiency workers represent a small share of the total labor force.

It has long been recognized that the impact of business cycle fluctuations on employment and total hours worked differs across subgroups of the population. Clark and Summers (1981) find that while teenagers comprise less than 10% of the U.S. population, they account for more than a fourth of cyclical employment fluctuations. Elsby, Hobijn, and Sahin (2010) find that younger, less educated workers, and individuals from ethnic minorities experienced steeper increases in joblessness during all of the last six U.S. recessions, mainly because of a larger fall in the exit rate from unemployment.

These observations suggest time-varying heterogeneity in the productivity level of the unemployed, as measured by observable characteristics, could have a strong impact on the volatility of aggregate labor market variables. This hypothesis has received only mixed support in empirical studies. Yet a worker's productivity may also depend on unobservable characteristics, and many theoretical frameworks imply that separations disproportionately hit low-productivity workers. In fact, a large literature has documented large wage differentials among observationally equivalent workers, and unexplained wage differentials are often found to be of the order of 70% of total wage inequality (Mortensen 2003).

Our assumption that workers with different efficiencies compete for the same position, and that firms need to meet workers to screen out less efficient candidates, is supported by evidence on the amount of resources spent by firms on recruitment. Villena-Roldan (2008) reports evidence from the National Employer Survey 1997 showing that firms interview a median of five applicants per vacancy and spend on average $4,200 on recruiting activities per recruited worker. A survey conducted by the Saratoga Institute (2000) reports that the direct cost of filling a single position—such as the costs of advertising, travel, but excluding the salaried time of managers conducting interviews—averages $4,588. Manning (2011) considers the empirical evidence on hiring costs, and finds that these costs are of the order of 5% of the total wage bill.[1]

1. Additional indirect evidence supporting our approach is found in the 2007 National Employers Skill Survey (Learning and Skill Council 2008), reporting that 21% of all vacancies among establishments in

While efficiency-heterogeneity in our model helps in explaining several observed stylized facts, such as higher unemployment volatility among low-efficiency workers, negative duration dependence of unemployment exit rates, and re-employment wage, a key contribution of our paper is to show under what conditions the composition effect is relevant.

Heterogeneity in efficiency levels by itself is not sufficient to generate amplification of productivity shocks on the unemployment rate. In an economy with large steady-state flows between employment and unemployment, a change in the employment level can be achieved with relatively small changes in separations and hiring. This implies the composition of the unemployed does not change much over the business cycle and the composition effect will contribute little to the volatility of unemployment. Thus, the larger labor flows that characterize the U.S. relative to many European economies imply a weaker composition effect. Pries and Rogerson (2005) report evidence that worker turnover is between two and three times larger in the U.S. than in Europe. Our results show that parameterizations consistent with labor data from the EU and the U.S. also lead to a much weaker composition effect in the U.S.

On the other hand, the impact of a fall in productivity that disproportionately affects low-efficiency workers relative to high-efficiency workers greatly amplifies the composition effect and the volatility of unemployment. A technology shock biased against low-efficiency workers can lower average productivity of the pool of unemployed workers, reduce vacancy posting, and lower the vacancy yield. In turn, this leads to a fall in the exit rate from unemployment for all workers, leading to higher economy-wide unemployment. If the notion of labor heterogeneity in our model is interpreted more broadly, it might help account for the empirical evidence for the recent U.S. experience which indicates the shortfall in employment has been especially hard in specific sectors (construction and manufacturing sectors) while vacancy yields have been below expectations across all the sectors during the recovery (Daly, Hobijn, and Valletta 2011).

We use our model to compare the effectiveness of alternative monetary policy rules. Following a fall in productivity, policy rules that are more expansionary are more effective at reducing the fall in output than the fall in employment. Moreover, stabilizing employment comes at a great cost in terms of inflation. Productivity shocks biased against low-efficiency workers pose even more of a conundrum to the policymaker, since the unemployment–inflation trade-off worsens and is very unfavorable, even in an economy with large average gross labor flows.

Our modeling framework is related to several contributions in the literature. 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, 2011, 2012), Walsh (2003, 2005), and Galí (2011). However,

England are considered hard to fill because of skill shortage. About a third of of these vacancies are hard to fill because of a lack of oral communication or customer-handling skills, rather than for lack of a specific technical skill.

these contributions with the exception of Walsh (2003, 2005) assume an exogenous separation rate, and all these previous papers assume homogenous workers.

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

Our approach is closer to the models of Rogerson and Pries (2005) and Pries (2008). In Rogerson and Pries, matches have persistent job-specific productivity, and firms screen for workers based on limited information on their productivity. As the match productivity is revealed over time, separations take place. Contrary to our approach, in Rogerson and Pries 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. Pries shows in a model with worker-specific productivity levels and exogenous separation rates that the composition effect has a large impact on the cyclical value of vacancies and thus on the behavior of employment flows. While our framework relies on a similar mechanism in affecting incentives to post vacancies, we relate the composition effect to the size of gross labor flows and the possibility of changes in TFP biased against low-efficiency workers. We also provide a framework with nominal rigidities that allows alternative monetary policies to be analyzed. In addition, Pries sets the relative covariance of separation rates for high- and low-productivity workers exogenously; this covariance is endogenous in our model and can vary depending on the nature of the shock processes.

The paper is organized as follows. Our model is presented in Section 1. The role of heterogeneity in efficiency levels and the composition effect is investigated in a calibrated version of the model in Section 2. This section also discusses evidence on cross-country differences in labor flows and on the impact of alternative policy rules for monetary policy. In Section 3, we review some of the empirical evidence on worker heterogeneity and labor market dynamics and discuss the consistency of this evidence with our model. Conclusions are discussed in the final section.

1. A MODEL WITH HETEROGENEITY IN EFFICIENCY LEVELS AND NONDIRECTED SEARCH

The model consists of households, wholesale and retail firms, and a monetary authority. Wholesale firms produce a homogenous good, which is sold in a competitive market to retail firms, of which there is a continuum of mass one. Retail firms

sell differentiated goods to households, and the retail sector is characterized by monopolistic competition and price stickiness as in standard New Keynesian models.

1.1 Overview of the Labor Market and Labor Flows

Workers are assumed to be heterogeneous with respect to their efficiency level; for simplicity, we assume workers are of two types, endowed either with a high (h) or low (l) efficiency level. Firms post vacancies to which unemployed workers apply. Firms must interview applicants to determine the worker's type. Thus, the job search and recruitment process involves both interviewing and screening. The aggregate number of interviews per period is determined through random matching as in standard matching models of the labor market. We assume all job seekers have identical interview-finding probability regardless of efficiency level. At the interview, the job applicant is screened. Not all interviews result in hires. We assume that if the efficiency level is revealed in the interview to be h, the worker is hired and produces with probability equal to one. That is, we assume the firm is able to identify a high-efficiency worker in the interview and the productivity of an h worker is high enough that it guarantees a positive surplus in all states.[2]

The productivity of low-efficiency workers is assumed to be stochastic. Each period, regardless of whether employed or unemployed, each low-efficiency worker i receives a new idiosyncratic stochastic productivity level $a_{i,t}$. 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.[3] While all high-efficiency unemployed workers who are interviewed are subsequently hired, only low-efficiency unemployed 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 be shown to 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 cutoff 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 idiosyncratic productivity shocks, all employed workers are subject to an aggregate productivity shock z_t. We also allow for asymmetric productivity shocks z_t^h, z_t^l. Hence, the total productivity of a low-efficiency worker-hour i at t is $z_t z_t^l a_{i,t}$ while that of a high-efficiency worker-hour is $z_t z_t^h$.

2. This assumption is for simplicity as it will imply that endogenous separations and interviews that do not lead to hires only involve low-efficiency workers.

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

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

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

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 has a low efficiency level. Let S^j be the number of type j workers who are seeking jobs, and let N^j be the number of type j workers who are employed. Then the probability a worker drawn from the pool of unemployed job seekers is low-efficiency is

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

while the share of employed workers of efficiency l is

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

The timing of activities 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.[4]

After exogenous separation occurs, all aggregate shocks realizations are observed. This allows firms to determine $\bar{a}_t$, the cutoff point for low-efficiency productivity that will determine hiring and retention. [5]

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 matching function. The probability a job seeker gets an interview is $k_t^w \equiv I_t/S_t$. Firms interview $k_t^f V_t$

4. 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.

5. We show that $\bar{a}_t$ is the same for all firms.

workers in the aggregate, where k_t^f is the probability a given vacancy receives an applicant to interview.

At time t idiosyncratic productivity shocks $a_{j,t}$ associated with employed low-efficiency workers and low-efficiency workers who are interviewed are observed. A fraction $1-\rho_t^n$ of type l workers receive productivity levels $a_{i,t} > \bar{a}_t$. So new hires H_t are given by the number of high-efficiency interviewees, all of whom are hired, plus the number of low-efficiency interviewees multiplied by the fraction of these with productivity levels that exceed $\bar{a}_t$:

$$H_t = (1-\gamma_t)k_t^w S_t + \left(1-\rho_t^n\right)\gamma_t k_t^w S_t = \left(1-\gamma_t\rho_t^n\right)k_t^w S_t.$$

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, has low efficiency but has a $a_{i,t} < \bar{a}_t$ and so is not hired) is $\gamma_t\rho_t^n$. In standard matching models, new hires equal $k_t^w S_t$. Screening implies new hires are less than this level and depend both on the endogenous average efficiency level of the pool of unemployed workers γ_t and on the aggregate productivity level, which we show below will affect 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

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

The total retention rate is $(1-\rho^x)(1-\xi_{t-1}\rho_t^n)$ and depends on the exogenous hazard ρ^x, the endogenous hazard for low-efficiency workers ρ_t^n, and the average efficiency level of beginning-of-period matches ξ_{t-1}. The share of low-efficiency employed workers evolves according to

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

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,

$$\gamma_t = \frac{L^l - (1-\rho^x)\xi_{t-1}N_{t-1}}{S_t}. \tag{2}$$

In deriving (1) and (2) 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.[6]

Since $a_{i,t}$ is i.i.d., the model does not generate a time-varying distribution of match-specific 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.

1.2 The Labor and Goods Markets

The wholesale sector. Wholesale firms post vacancies, interview and screen applicants, make hiring and retention decisions, and produce a homogenous output. Let h_t^h denote hours worked by an employed high-efficiency worker, and let $h_{i,t}^l$ be hours worked by employed low-efficiency worker i. All type h workers will work the same hours since they have the same productivity, but the hours of low-efficiency workers will depend on their idiosyncratic productivity realizations. 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 $\bar{a}_t$):

$$\begin{aligned} Q_t &= z_t z_t^l N_t^l \left[\frac{\int_{\bar{a}_t}^1 a_{i,t} h_{i,t}^l dF(a_i)}{1 - F(\bar{a}_t)} \right] + z_t z_t^h h_t^h N_t^h \\ &= \left\{ z_t^l \xi_t \left[\frac{\int_{\bar{a}_t}^1 a_{i,t} h_{i,t}^l dF(a_i)}{1 - F(\bar{a}_t)} \right] + (1 - \xi_t) z_t^h h_t^h \right\} z_t N_t, \end{aligned} \tag{3}$$

where $z_t z_t^j$ is aggregate productivity for all workers of efficiency level $j = [l, h]$ and $F(a)$ is the cumulative distribution function of the idiosyncratic productivity shocks. We assume the productivity of a match depends on a common productivity disturbance z_t, with the productivity z_t^l of l workers equal to z_t, and the productivity of h workers equal to $z_t^h = z^h z_t$. The constants z^h and z^l are used to parameterize the relative average productivity of l and h workers. Since $F(\bar{a})$ is the probability $a_{i,t} \leq \bar{a}_t$, $F(\bar{a}) = \rho_t^n$ is also the endogenous separation and screening rate.

The homogenous output of wholesale firms is sold to retail firms in a competitive goods market. The price of the wholesale goods is P_t^w; the aggregate price index for retail goods is P_t. We define $\mu_t = P_t / P_t^w$ as the retail-price markup.

6. Combining equations (1) and (2), it can be seen that job seekers at t who are of quality l arise from three sources: low-efficiency workers who were searching for jobs in $t - 1$ and failed to be hired, those employed 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.

Expressed in terms of final retail goods, the current surplus of a firm–worker match involving a high-efficiency worker is

$$s_t^h = \left(\frac{z_t z_t^h h_t^h}{\mu_t}\right) - \frac{v\left(h_t^h\right)}{\lambda_t} - w_t^{u,h} + q_t^h, \tag{4}$$

where h_t^h is chosen optimally to maximize the match surplus, $v(h_t^h)$ 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 have the same productivity, work the same number of hours, and generate the same surplus.

The surplus of a match involving a low-efficiency worker is

$$s_{i,t}^l = \left(\frac{a_{i,t} z_t z_t^l h_{i,t}^l}{\mu_t}\right) - \frac{v\left(h_{i,t}^l\right)}{\lambda_t} - w_t^{u,l} + q_t^l. \tag{5}$$

This differs from the expression for high-efficiency worker–firm matches because of the idiosyncratic productivity disturbance and the nondegenerate distribution of hours worked among low-efficiency workers. As is common in the literature on unemployment, we assume complete consumption risk sharing, so λ_t is the same for all workers.

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. Let $f(a_i)$ be the density function for $a_{i,t}$. The continuation values are therefore given by

$$q_t^h = \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t}\right) \left[(1-\rho^x) s_{t+1}^h + w_{t+1}^{u,h}\right] \tag{6}$$

and

$$\begin{aligned} q_t^l &= \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t}\right) \left[(1-\rho^x)\left(1-\rho_{t+1}^n\right) E_t\left(s_{i,t+1}^l \middle| a_{i,t} > \bar{a}_{i,t}\right) + w_{t+1}^{u,l}\right] \\ &= \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t}\right) \left[(1-\rho^x) \int_{\bar{a}_{t+1}}^1 s_{i,t+1}^l f(a_i) da_i + w_{t+1}^{u,l}\right]. \end{aligned} \tag{7}$$

To determine $w_t^{u,j}$, we assume that w^u is the value of time spent unemployed (home production or an unemployment benefit) and that 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

$$w_t^{u,h} = w^u + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t}\right) \left(k_{t+1}^w \eta s_{t+1}^h + w_{t+1}^{u,h}\right), \tag{8}$$

while for a low-efficiency worker it is

$$\begin{aligned} w_t^{u,l} &= w^u + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[k_{t+1}^w \eta (1 - \rho_{t+1}^n) E_t \left(s_{i,t+1}^l | a_{i,t} > \bar{a}_{i,t} \right) + w_{t+1}^{u,l} \right] \\ &= w^u + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[k_{t+1}^w \eta \int_{\bar{a}_t+1}^1 s_{i,t+1}^l f(a) da + w_{t+1}^{u,l} \right]. \end{aligned} \tag{9}$$

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

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

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

$$v_h \left(\hat{h}_t^l \right) \equiv \frac{\partial v \left(\hat{h}_t^l \right)}{\partial \hat{h}_t^l} = \left(\frac{\bar{a}_t z_t z_t^l}{\mu_t} \right) \lambda_t.$$

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

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

$$\rho_t^n = F(\bar{a}_t).$$

This is also the probability a low-efficiency worker who receives an interview is not hired. If the aggregate productivity shock is low, $\bar{a}_t$ 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 as they are more likely to be screened out during the interview process.

Hours. Hours maximize the joint surplus in a match s_t^h, $s_{i,t}^l$. For a high-efficiency worker, this implies

$$v_h \left(h_t^h \right) = \left(\frac{z_t z_t^h}{\mu_t} \right) \lambda_t. \tag{10}$$

For a low-efficiency worker of productivity $a_{i,t}$, this implies

$$v_h(h^l) = \left(\frac{a_{i,t} z_t z_t^l}{\mu_t}\right) \lambda_t;\ a_{i,t} \geq \bar{a}_t. \tag{11}$$

Vacancies. 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

$$k_t^f (1-\eta)\left[(1-\gamma_t)s_t^h + \gamma_t \int_{\bar{a}_t}^1 s_{i,t}^l f(a_i) da_i\right] = \kappa, \tag{12}$$

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.

Given the pool of job seekers S_t and the number of vacancies V_t posted by firms, the number of new interviews is determined by a standard matching function $m(S_t, V_t)$. This is taken to be Cobb–Douglas with constant returns to scale:[7]

$$m(S_t, V_t) = \psi S_t^{\alpha} V_t^{1-\alpha} = \psi \theta_t^{1-\alpha} S_t, \qquad 0 < \alpha < 1, \quad \psi > 0, \tag{13}$$

where $\theta_t \equiv V_t / S_t$ is the standard measure of labor market tightness. Because of worker heterogeneity, the probabilities of being interviewed and being hired will differ by the worker's efficiency level. The probability an unemployed worker obtains an interview, k_t^w, is

$$k_t^w = \frac{m(S_t, V_t)}{S_t} = \psi \theta_t^{1-\alpha}. \tag{14}$$

This is the same for all job seekers. Similarly, the probability a firm with a posted vacancy finds an applicant, k_t^f, is

$$k_t^f = \frac{m(S_t, V_t)}{V_t} = \psi \theta_t^{-\alpha}. \tag{15}$$

7. Constant returns to scale is consistent with the empirical evidence when applied to new hires; see Petrongolo and Pissarides (2001).

Compared to the standard homogeneous workers setup, k_t^w is the probability a firm obtains an interview, and k_t^f is the probability an interview slot will not go unfilled. The job-finding probability is identical to the interview rate for high-efficiency workers, 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. The overall job-finding probability can be defined as $\gamma_t k_t^{w,l} + (1 - \gamma_t) k_t^w$. With worker-specific productivity, a job opening that would be filled and lead to production if a high-efficiency applicant is interviewed may go unfilled if a low-efficiency worker is interviewed.

1.3 Households

The representative household purchases consumption goods, holds bonds, and supplies labor. Since some workers will be matched while others will not be, and workers differ in their productivity and hours worked, distributional issues arise. To avoid these issues, we follow the literature in assuming household's pool consumption by viewing the household as consisting of a continuum of members of various efficiency levels, some of whom will be employed, others unemployed.[8] Households are also the owners of all firms in the economy.

Households maximize

$$\mathrm{E}_t \sum_{i=0}^{\infty} \beta^i \left[\frac{\mathcal{C}_{t+i}^{1-\sigma}}{1-\sigma} - v\left(h_{t+i}^h\right)(1 - \xi_{t+i}) N_{t+i} - \xi_{t+i} N_{t+i} \int_{\bar{a}_t}^1 v\left(h_{i,t+i}^l\right) f(a) da \right], \quad (16)$$

where $\sigma > 0$ is the coefficient of relative risk aversion, $\mathcal{C}_t$ is the sum of a market-purchased composite consumption good C_t and home-produced consumption by unemployed workers $C_t^u = (1 - N_t) w^u$.

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

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

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

$$c_{kt} = \left(\frac{p_{kt}}{P_t} \right)^{-\theta} C_t, \quad (17)$$

8. This assumption is common in search and matching models of the labor market (see, e.g., den Haan, Ramey, and Watson 2000).

where the aggregate retail price index P_t is defined as

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

In (16), the term

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

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 $v(h_{t+i}) = \ell h_{t+i}^{1+\chi}/(1+\chi)$.

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

$$\lambda_t = \beta(1+i_t)\mathrm{E}_t\left(\frac{P_t}{P_{t+1}}\right)\lambda_{t+1}, \tag{18}$$

where λ_t is marginal utility of total consumption at time t.

1.4 Retail Firms

Each retail firm purchases wholesale output, which it then converts into a differentiated final good that is sold to households and wholesale firms. Retail firms maximize profits subject to a constant returns to scale technology for converting wholesale goods into final goods, the demand functions (17), and a restriction on the frequency with which they can adjust their price.

Retail firms adjust prices according to the Calvo updating model. 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 goods relative to the price of final output, P_t^w/P_t. This is just the inverse of the markup of retail over wholesale goods.

A retail firm k that can adjust its price in period t chooses p_{kt} to maximize

$$\sum_{s=0}^{\infty}(\omega\beta)^s\,\mathrm{E}_t\left[\left(\frac{\lambda_{t+s}}{\lambda_t}\right)\left(\frac{p_{kt}-P_{t+s}^w}{P_{t+s}}\right)Y_{t+s}(k)\right],$$

subject to

$$Y_{t+s}(k) = Y_{t+s}^d(k) = \left[\frac{p_{kt}}{P_{t+s}}\right]^{-\varepsilon} Y_{t+s}^d, \tag{19}$$

where Y_t^d is aggregate demand for the basket of final goods. The first-order condition for those firms adjusting their price in period t is

$$p_{kt}\mathrm{E}_t \sum_{s=0}^{\infty}(\omega\beta)^s \left(\frac{\lambda_{t+s}}{\lambda_t}\right)\left[\frac{p_{kt}}{P_{t+s}}\right]^{1-\varepsilon} Y_{t+s}$$

$$= \left(\frac{\varepsilon}{\varepsilon-1}\right)\mathrm{E}_t \sum_{s=0}^{\infty}(\omega\beta)^s \left(\frac{\lambda_{t+s}}{\lambda_t}\right)\left(\frac{1}{\mu_{t+s}}\right)\left[\frac{p_{kt}}{P_{t+s}}\right]^{1-\varepsilon} Y_{t+s}.$$

When linearized around a zero-inflation steady state, these conditions yield a New Keynesian Phillips curve in which the retail price markup $\mu_t \equiv P_t/P_t^w$ is the driving force for inflation. As in a standard Phillips curve, the elasticity of inflation with respect to real marginal costs will be $\delta \equiv (1-\omega)(1-\beta\omega)/\omega$.

1.5 Monetary Policy

We assume that the monetary authority in this economy implements monetary policy using the nominal interest rate as the policy instrument. Given our assumptions on price updating, the monetary authority can implement the flexible price allocation with constant markups using a policy of complete price stability, although this policy does not necessarily deliver the first-best allocation because of the search frictions in the labor market (see Ravenna and Walsh 2012).

1.6 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

$$Y_t = C_t + \kappa V_t. \tag{20}$$

2. THE IMPACT OF EFFICIENCY HETEROGENEITY ON UNEMPLOYMENT DYNAMICS

This section evaluates the role the composition effect plays in contributing to the response of unemployment to productivity shocks. We analyze the dynamic response of the economy in response to two kinds of shocks: a negative aggregate TFP shock and an asymmetric shock that affects disproportionately the TFP of low-efficiency workers. We take as a benchmark an economy where the monetary authority enforces price stability. This case shows how the economy would respond if prices were flexible and nominal rigidity did not result in a binding constraint for firms. We then

consider the case of alternative monetary policy rules, where firms' markups are time varying and the equilibrium can deviate from the flexible-price allocation.

The impact of the change in the composition of the unemployment pool on the unemployment rate in a recession works through two channels: first, by changing the relative quantity of low- to high-efficiency workers searching for a match (the direct composition effect), and second, by changing the incentive of firms and applicants to form matches (the indirect incentive effect). As the matches with the least productive workers separate in a downturn, their share in the unemployment pool increases. As a consequence, the average productivity of the unemployed falls by more than the average productivity of the labor force, and the outflow rate from unemployment decreases by more than it would in a model with homogeneous workers.

The direct composition effect can be illustrated through the equation defining the unconditional outflow rate. For a randomly chosen unemployed worker, the job-finding probability is the weighted average of the job-finding probability for l and h workers:

$$k_t^{\text{job},w} = \gamma_t k_t^w \Pr\left(s_{i,t}^l > 0\right) + (1-\gamma_t)k_t^w. \tag{21}$$

The probability of finding a job for an l worker depends on the interviewing rate k_t^w and on the probability that the idiosyncratic productivity shock yields a positive match surplus. Both will fall in a recession; thus, the job-finding probability falls by more for the l workers than for the h workers. With heterogeneous efficiency levels, the larger the increases in the share γ_t of l workers, the larger the amplification in the fall of the unconditional job-finding probability.

The direct composition effect also works by lowering the probability of filling a vacancy for firms. Firms become more selective in a recession. For a given number of posted vacancies, an increase in γ_t implies the chance that a randomly interviewed worker will be hired is lower. The impact of γ_t on the hiring probability can be described by the screening rate, that is, the unconditional rate at which an interviewee is screened out:

$$scr_t = \gamma_t \rho_t^n = \gamma_t\left[1 - \Pr\left(s_{i,t}^l > 0\right)\right]. \tag{22}$$

Ceteris paribus, in a recession the screening rate increases for two reasons. First, as in any search model of the labor market with endogenous separation, the separation rate ρ_t^n increases. Second, the likelihood that an interviewee is a low-efficiency worker γ_t also increases.

The indirect incentive effect of the change in the composition of the unemployment pool occurs through the change in the value of a vacancy. Equations (12) and (15) imply the vacancy posting condition can be written as:

$$\frac{V_t}{S_t} = \left\{\frac{\psi}{\kappa}(1-\eta)\left[\gamma_t \int_{\bar{a}_t}^1 s_{i,t}^l f(a_i)da_i + (1-\gamma_t)s_t^h\right]\right\}^{1/\alpha}. \tag{23}$$

The right-hand side of (23) depends on the expected surplus from a match. Since the surplus s_t^h from a high-efficiency worker is higher than the expected surplus $\int_{\bar{a}_t}^1 s_{i,t}^l f(a_i) da_i$ 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 the fall in the number of vacancies per searching worker, and the larger the fall in the interview rate k_t^w, which is proportional to V_t/S_t through equation (14).

In summary, any shock that results in an increase in the low-efficiency unemployed share worsens the probability of exiting unemployment for *all* workers, and of filling a position for firms. From the perspective of the job applicants, the chance of exiting unemployment falls since fewer vacancies are posted. From the perspective of the firm, the average worker has a higher likelihood of being drawn from the low-efficiency pool and when interviewed, low-efficiency workers have a lower likelihood of being hired.

2.1 Parameterization

To evaluate the dynamic behavior of the model economy, we adopt a baseline parameterization based on European data. The model is very parsimonious and contains a limited number of parameters. 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 $z_{ss}^h/(z_{ss}^l \int_0^1 a_i dF(a_i))$, 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 the average quarterly rate for the EU15 group of countries, over the sample 1993:1 to 2010:4. 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 match the l-efficiency workers' unemployment rate with the rate for the 16–24 age group, and the h-efficiency unemployment rate with the rate for the 25–74 age group, reported in the Labor Force Survey compiled by Eurostat. The steady-state hours per worker h_{ss}^{av} and the probability of a match between an applicant and a vacancy k_{ss}^f are parameterized to standard values in the labor search literature. The share of output devoted to hiring activities is in line with empirical evidence reported in Ravenna and Walsh (2008).

The steady-state aggregate separation rate ρ^x is set according to available average separation data (Blanchard and Galí 2010). Our parameterization strategy takes as given a value for the exogenous separation rate, but the aggregate separation rate ρ_{ss} turns out to be close in value to the exogenous rate. The choice for the

TABLE 1

BASELINE PARAMETERIZATION

		Steady-state values
Unemployment rate	u_{ss}	8.7%
Unemployment rate: *low-efficiency* labor	u^l_{ss}	17.7%
Unemployment rate: *high-efficiency* labor	u^h_{ss}	7.4%
Average hours per worker	h^{av}_{ss}	0.25
Vacancy posting cost share of output	$\frac{\kappa V_{ss}}{Y_{ss}}$	0.05
Probability of vacancy matched with applicant	k^f_{ss}	0.7
Parameters		
Vacancy elasticity of matches	α	0.6
Discount factor	β	0.99
Inverse of labor hours supply elasticity	χ	2.5
Relative risk aversion	σ	1
Steady-state inflation rate	π_{ss}	1
Workers' share of surplus	η	0.35
Exogenous separation rate	ρ^x	3.4%
Implied steady-state separation rate	ρ_{ss}	3.8%
AR(1) parameter for technology shock z_t	ρ_z	0.95
Price elasticity of retail goods demand	θ	6
Average retail price duration (quarters)	$\frac{1}{1-\omega}$	3.33
Steady-state markup	μ	1

NOTES: Baseline parameterization based on EU15 data. Unemployment rate for low- and high-efficiency workers is given, respectively, by the rate for the 16–24 and 25–74 age group. Sample: quarterly data, 1993:1–2010:4.
SOURCE: Eurostat (2011).

remaining parameters follows the recent literature on business cycle models with search unemployment and nominal rigidities.[9]

The parameterization implies that $\bar{\gamma}$, the share of l workers in the labor force L, is 13.4%. Because the separation rate of l workers is about twice as large as the overall separation rate, their share γ_{ss} in the pool of job seekers is 22.6%, while their share ξ_{ss} in the employment pool is 11.8%. Thus, low-efficiency workers are overrepresented in the pool of unemployed.

To illustrate the relevance of the size of average labor flows, we compare our baseline parameterization to two alternative economies. The first alternative has the same steady-state level of output and unemployment as our (EU) baseline but larger steady-state flows. The second alternative, corresponding to a U.S. parameterization and discussed in Section 2.3, has larger labor flows but a smaller steady-state unemployment level.

In the first, large labor flow alternative parameterization, we assume the economy draws on a labor force where low-efficiency workers are less productive, relative to the baseline, yielding a larger steady-state separation rate. Table 2 shows that in this economy, the average productivity of high relative to low-efficiency workers is

9. The only exception is given by η, the workers' share of surplus, which given our choice of α implies the Hosios condition is not met. The value of $\eta = 0.35$ was chosen to be as close as possible to the Hosios condition, while ensuring determinacy of the equilibrium.

TABLE 2

ALTERNATIVE PARAMETERIZATIONS FOR DIFFERENT SIZES OF STEADY-STATE LABOR FLOWS

	Baseline	Large labor flows
Parameters		
Average productivity of high-efficiency workers	0.76	0.76
Average productivity of low-efficiency workers	0.5	0.40
Relative productivity of high-/low-efficiency workers	1.53	1.90
Average productivity of labor force	0.73	0.71
Matching function productivity ψ	0.42	0.80
Vacancy posting cost κ	0.16	0.05
Steady state		
Overall separation rate ρ_{ss}	0.038	0.062
Endogenous separation rate ρ^n_{ss}	0.039	0.51
Low-efficiency unemployment share γ_{ss}	0.23	0.66
Low-efficiency employment share ξ_{ss}	0.12	0.06
Unemployment duration (quarters)	3.36	2.10

NOTES: 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.

higher, while the average productivity of the labor force is very similar. To obtain an alternative economy with identical unemployment rate and output as the baseline, we also assume a higher productivity ψ of the matching function and a lower vacancy-posting cost κ. In this way, the steady-state outflow from unemployment is large enough to balance the higher steady-state inflow at the same level of steady-state unemployment.

When the relative productivity of high- to low-efficiency workers increases, the endogenous separation rate ρ^n_{ss} for low-efficiency workers increases to 51%, while it is only 3.9% in the baseline parameterization. The overall separation rate increases by a smaller amount, since the share of low-efficiency workers in the labor force is unchanged relative to the baseline, and equal to 13.4%. The increase in the separation rate implies the share of low-efficiency workers in the unemployed pool rises from 23% in the baseline to 66% in the alternative parameterization. This implies that in a recession the percent increase in the separation rate, and in the share of low-efficiency unemployed, required to achieve the equilibrium change in employment is smaller relative to the baseline parameterization. Finally, the larger size of gross labor flows implies that, while the unemployment rate is identical across the two economies, the unemployment duration is about 60%longer in the baseline parameterization.

2.2 Gross Labor Flows and the Relevance of the Composition Effect

Our first experiment considers the impact of a persistent fall in aggregate TFP, comparing the economies with small and large gross labor flows. This comparison is useful in assessing the impact of the composition effect, since in the large labor flows economy the efficiency composition of unemployment is essentially unchanged by an aggregate TFP shock.

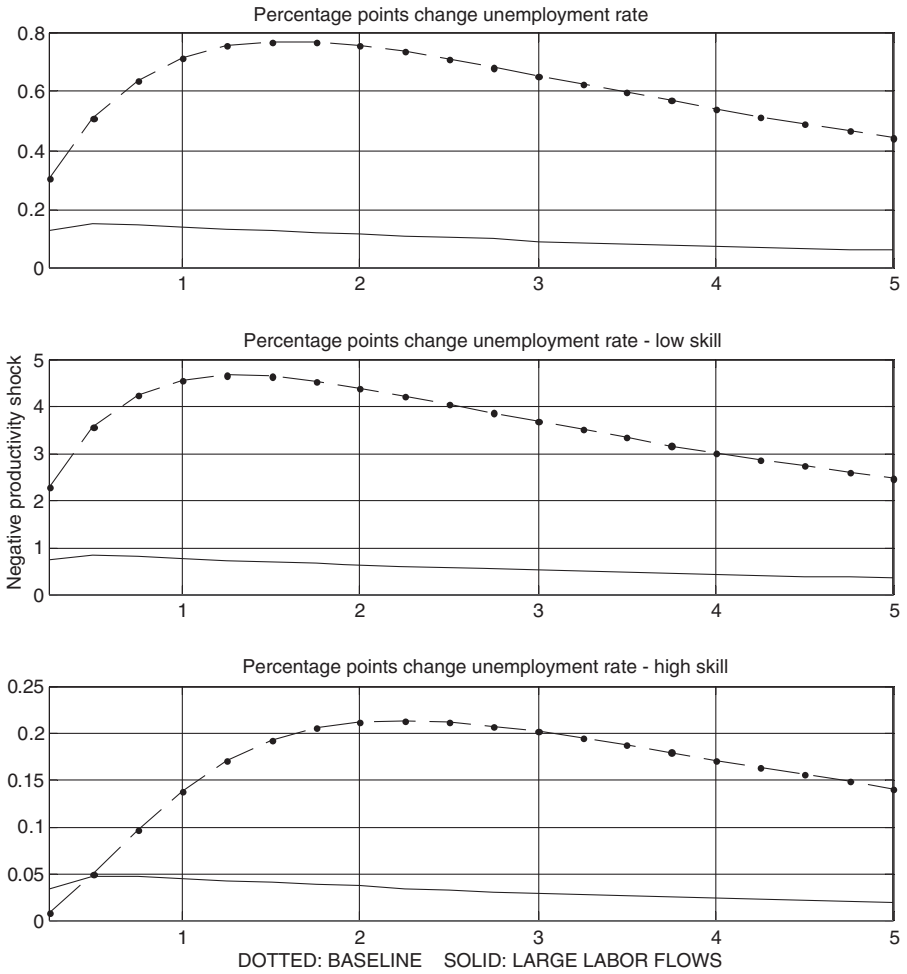

FIG 1. Impulse Response to a 1% Negative TFP Shock z_t under Price Stability for the Baseline Parameterization, and the Steady-State Large Labor Flows Parameterization Described in Table 2.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Change in unemployment rate for total, low- 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.

Figure 1 shows the dynamic response on the aggregate unemployment rate and the unemployment rates for the two types of workers when monetary policy maintains price stability. The change in the unemployment rate for the low-efficiency workers is over 20 times as large as for the high-efficiency workers, since low-efficiency workers experience a larger fall in job-finding probability and a larger increase in unemployment duration. The effect on the overall unemployment rate is relatively small in the parameterization with large labor flows, a feature that is common to search models of the labor market with Nash bargaining. In the baseline parameterization,

the impact of the TFP shock is significantly amplified. The unconditional volatility of employment relative to output σ_n/σ_y is equal to 0.65 in the baseline parameterization and only 0.14 in the alternative one. Note that this amplification is obtained with a steady-state share of low-efficiency workers in the employment pool of only 11.8% and in the labor force of only 13.4%.

An implication of a strong composition effect is a considerable delay in the response of unemployment to a fall in productivity and its subsequent sluggish recovery. The peak response in overall unemployment occurs after seven quarters in the baseline case, and two quarters in the alternative one. The lag depends on the progressive build-up of a larger share of low-efficiency workers in the unemployed pool (who have a lower outflow rate from unemployment) and the reduction in the incentive to post vacancies.

Figure 2 shows the log-deviation of selected variables in response to a negative productivity shock. On impact, the fall in output is very similar across the two parameterizations. When the composition effect is at work, though, output keeps falling for the first five quarters, and after 3 years production is still below trend by an additional 0.4% compared to the alternative parameterization. The increase in the separation rate—driven entirely by the firing of low-efficiency workers—raises the share of less productive workers in the unemployment pool by over 15% in the baseline economy. Since the composition effect amplifies the fall in the average productivity of the jobless, the unconditional job-finding probability falls sharply. In the alternative parameterization, the response of the separation rate is muted; thus, the composition of employment shifts in favor of h workers, but the efficiency composition of the unemployment pool hardly changes. This implies that the average fall in productivity among the unemployed is nearly identical to the fall in aggregate TFP.

To single out the role of the composition effect in reducing the flow out of unemployment, Figure 3 compares the behavior of different variables to the counterfactual built from (23) under the assumption that γ_t remains constant. The first panel of Figure 3 shows that as the average efficiency level of the pool of unemployed worsens, the fall in productivity among the unemployed more than doubles relative to an economy with homogeneous workers. Moreover, since the excess of low-efficiency workers relative to steady state accumulates slowly in the unemployment pool, the fall in TFP for the average unemployed worker peaks nearly a year and half later than aggregate TFP.

Heterogeneity in the efficiency level amplifies unemployment volatility because the fall in productivity of the overall unemployment pool is much more severe than for the labor force overall or for those workers who remain employed. The fall in TFP among the unemployed amplifies the volatility of employment relative to output, since the productivity of the employed workers—the majority in the economy—has not changed. In our baseline parameterization, a 5% increase in the low-efficiency unemployment share corresponds to about a 1 percentage point increase of the low-efficiency unemployment share, from 22.6% to 23.6% Since the extra percentage point of low-efficiency workers has replaced high-efficiency workers whose productivity is 50% higher, TFP among the unemployed falls by roughly

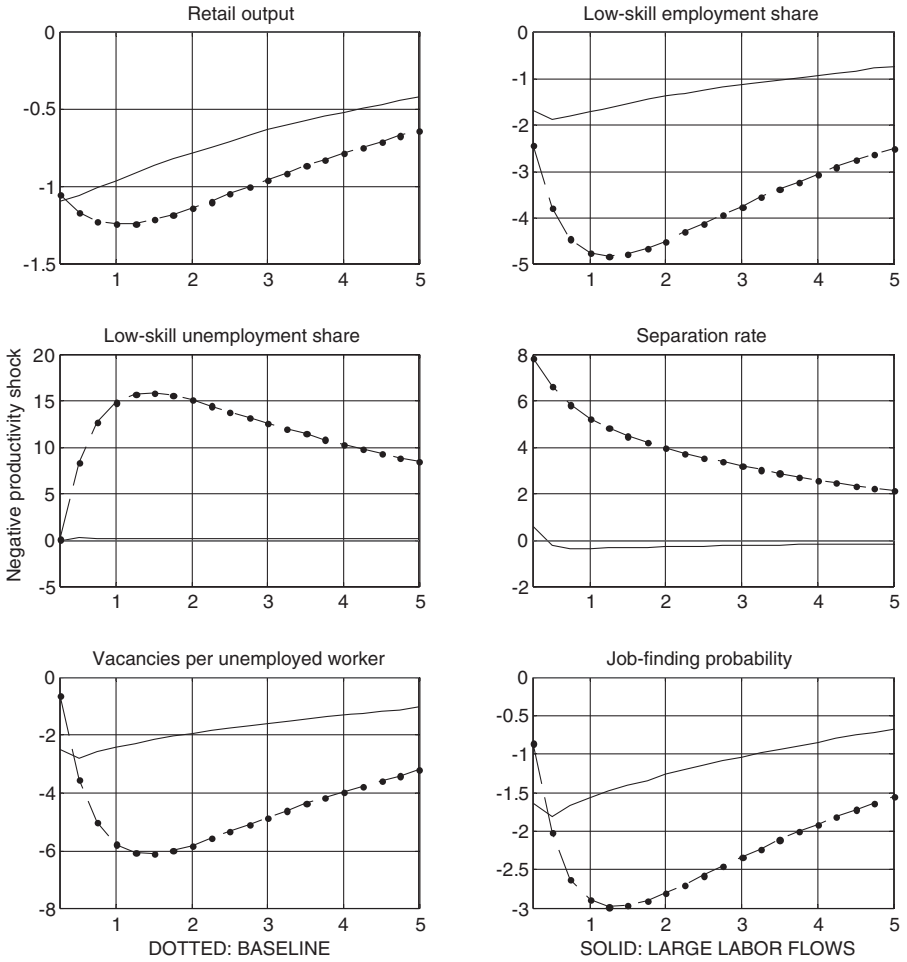

FIG 2. Impulse Response to a 1% Negative TFP shock z_t Under Price Stability for the Baseline Parameterization and the Steady-State Large Labor Flows Parameterization Described in Table 2.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Scaling in percent; horizontal axis in years.

0.5%. Note that while our parameterization implies that low-efficiency workers are substantially less productive than high-efficiency workers, they represent only 13.4% of the labor force. Thus, in steady state the average TFP of the employed worker-hour is only 4.5% higher than the average TFP for the unemployed.[10]

10. Residual wage inequality, which can be interpreted as reflecting unmeasured differences in productivity, has been documented to be very large. Hornstein, Krusell, and Violante (2007) use 1990 U.S. Census data to show that the ratio of the mean wage to the 10th percentile is 1.83 even conditioning on low-skill occupations and a set of workers with less than 10 years of experience.

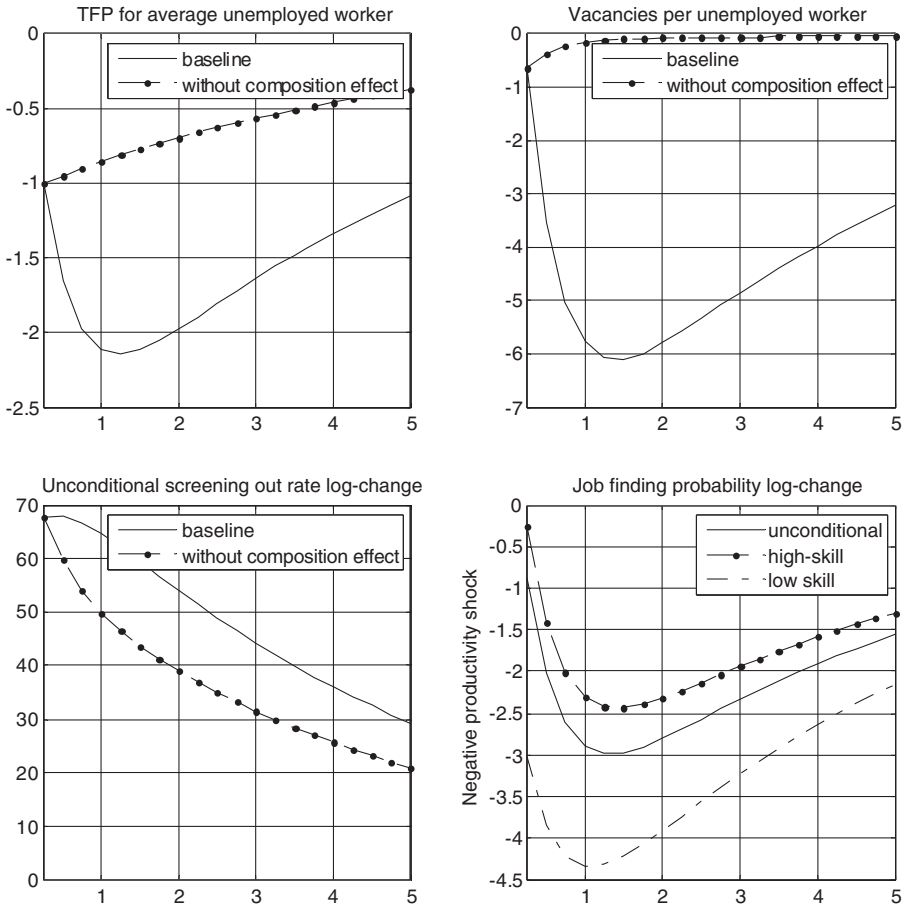

FIG 3. Impulse Response to a 1%Negative TFP Shock z_t Under Price Stability for the Baseline Parameterization (Table 1).

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Scaling in percent. Impulse responses without composition effect assume share of low-efficiency unemployed is constant at $\gamma_t = \gamma_{ss}$. Horizontal axis in years.

The top right panel of Figure 3 compares the behavior of the log-change in vacancies per unemployed worker to the counterfactual in which γ_t remains constant. Virtually all the fall in the incentive to post vacancies originates from the change in the efficiency composition of the unemployed. Additionally, the composition effect increases the likelihood that any firm that posts a vacancy will end up interviewing a low-efficiency worker, so the probability an interview actually results in a hire decreases as more interviewees will be screened out. The lower left panel of Figure 3 compares the screening rate defined in (22) to the counterfactual assuming γ_t is constant. The composition effect increases the screening rate by up to 40%.

TABLE 3

PERCENT CONTRIBUTION TO LOG-CHANGE IN JOB FINDING RATE

	ρ_t^n	γ_t	k_t^w
Baseline	15%	5%	80%
Large labor flows	31%	4%	65%

NOTES: Contribution to cumulative log-change of unconditional job-finding rate $k_t^{\text{job},w}$ over 20 quarters following a TFP shock, of the endogenous separation rate ρ_t^n, the low-efficiency unemployment share γ_t, and the probability k_t^w of an unemployed worker obtaining an interview with a firm.

The bottom right panel of Figure 3 shows the behavior of the unconditional outflow rates for low- and high-efficiency workers. The unconditional rate falls in part because both k_t^w and $k_t^{w,l}$ fall, but it also falls because the weight on $k_t^{w,l}$ increases in the overall average job-finding rate.

Table 3 exploits the log-linear approximation to (21) to compute the contribution to the change in the outflow rate originating from the change in the separation rate for new matches ρ_t^n, the share of low-efficiency unemployed γ_t (the direct composition effect), and the probability k_t^w that a firm matches a vacancy with an interview (the indirect incentive effect). When the composition effect is at work, the change in the job-finding rate that is driven by the increase in the separation rate for new matches falls by half relative to the economy without a composition effect, from 31% to 15%. Most of the difference is explained by the larger fall in the probability of an interview taking place.[11]

In summary, in an economy with large steady-state labor flows between employment and unemployment, a change in the employment level can be achieved with a relatively small change in separations and hiring. This implies that the efficiency composition of the unemployment pool does not change much in a business cycle, the change in productivity among unemployed workers is not amplified, and neither is the outflow rate from unemployment. An economy with smaller gross labor flows—even with an identical unemployment rate in steady state—will experience a sharper increase in separations during a recession, a significant worsening of the unemployment pool efficiency level, and a larger fall in the outflow rate from unemployment. This ultimately leads to a slower recovery, as the efficiency composition of the unemployment pool slowly reverts to its steady state.

2.3 A Comparison of the EU and U.S. Parameterizations

In this section we compare the impact of a productivity shock for the baseline parameterization, obtained using data for the EU15 group of countries, and a parameterization based on U.S. data.

11. The change in γ_t plays a similar, and small, role in the fall in the job-finding probability. This depends on the fact that the share of low-efficiency workers is small among the unemployed in the baseline parameterization, while it is large but unresponsive to the productivity shock, in the alternative parameterization.

A very extensive literature has documented the differences in labor flows between the U.S. and large European economies. Elsby, Hobijn, and Sahin (2008) find that unemployment inflow and outflow rates are positively correlated, with continental European countries characterized by low rates of both inflow and outflow. The average of the inflow and outflow rates in France, Germany, Italy, Portugal, and Spain ranged from 4.8% (Italy) to 10.2% (Spain). By way of contrast, the rate averaged 40% in the U.S. The estimated rate of outflow from unemployment for Spain was 1% while rates for France, Germany, Italy, and Portugal were even lower. For the U.S., the comparable figure was estimated to be 3.6%. Elsby, Hobijn, and Sahin (2008) argue that inflows contribute only about 20% of the time series variation of unemployment rates in Anglo-Saxon and Nordic countries, a finding consistent with Shimer (2005). However, the corresponding figure for continental European economies is 50%, suggesting a much larger relative role is played by variations in the inflow to unemployment in accounting for fluctuations in European unemployment.

Jung and Kuhn (2011) provide additional evidence on cross-country differences in employment dynamics using U.S. and German data. While the average German job-finding and firing rate are, respectively, one-fifth and one-fourth of the U.S. one, the firing rate is 2.5 times more volatile in Germany than in the U.S. Firing contributes between 60% and 70% to the German unemployment volatility. The authors report evidence supportive of an important role for heterogeneity in workers' efficiency level: 75% of all firings happen for matches with tenure less than 2 years, and the majority of jobs destroyed fall at the lower end of the earnings distribution.[12]

In our U.S. parameterization, the unemployment steady-state values are obtained averaging U.S. Bureau of Labor Statistics (BLS) quarterly data over 1948:1 to 2010:1. We identify unemployment rates for low- and high-efficiency workers with rates for workers' age groups 16–24 and over 24. While the U.S. has lower unemployment rates across all groups, the ratio of the type-specific to the aggregate unemployment rates is similar to the EU case. Table 4 shows the two sets of steady-state values matched under the two parameterizations. The steady-state aggregate separation rate is about twice as large in the U.S. parameterization, consistent with available average separation rate data (Shimer 2005).

The parameterization implies that the steady-state share of l workers in the labor force is 16% in the U.S. and 13.4% in the EU. The share of l workers in the pool of job seekers is similar across the two parameterizations, equal to 22.6% for the EUand 22.8% for the U.S. parameterization. Unemployment duration is half as long in the U.S. case, where it is equal to 1.71 quarters, relative to the EU case, where it is equal to 3.36 quarters.

12. The important role played by fluctuations in the rate of inflow into unemployment in European economies is inconsistent with the standard assumption of most recent monetary models with search and matching in the labor market, which typically assume a constant and exogenous separation rate (e.g., Gertler, Sala, and Trigari 2008, Ravenna and Walsh 2008, 2011, 2012, Gertler and Trigari 2009, Blanchard and Galí 2010).

TABLE 4

PARAMETERIZATION FOR EU AND THE U.S.

	Steady-state values	U.S.	EU
Unemployment rate	u_{ss}	5.7%	8.7%
Unemployment rate: *low-efficiency*	u^l_{ss}	11.6%	17.7%
Unemployment rate: *high-efficiency*	u^h_{ss}	4.4%	7.4%
Average hours per worker	h^{av}_{ss}	0.33	0.25
Exogenous separation rate	ρ^x	6.8%	3.4%
Implied steady-state separation rate	ρ_{ss}	7.4%	3.8%

NOTES: U.S. unemployment rate for low- and high-efficiency workers is given, respectively, by the rate for the 16–24 and over-24 age group, over the 1948:1–2010:1 sample (BLS 2011). EU unemployment rate for low- and high-efficiency workers is given, respectively, by the rate for the 16–24 and 25–74 age group for the EU15 countries, over the 1993:1–2010:4 sample (Eurostat 2011).

Gross labor flows are larger in the U.S. case. We parameterized the model so that the higher separation rate is primarily the result of a higher rate of exogenous separations, consistently with empirical evidence showing that the volatility of unemployment in the U.S. is largely explained by volatility in the outflow rate from unemployment. Thus, the implied endogenous separation rate is similar across parameterizations (ρ^n_{ss} is equal to 3.9% for the EU and 4.6% for the U.S.). Despite the fact that the difference in average labor flows across the two parameterizations originates from exogenous rather than endogenous separations (and thus also affects high-efficiency workers, contrary to our earlier experiment comparing small- and large-labor flows economies), the composition effect still turns out to be much smaller for the U.S. case. Figures 4 and 5 show that the impact of a 1% fall in TFP, reducing output by approximately 1% on impact across the two economies. The rise in unemployment in the U.S. case is less than half as large and peaks earlier than for the EU parameterization. The unemployment rate among low-efficiency workers increases by over 20 times the high-efficiency one in the EU case, and only by about 10 times in the U.S. case. If we identify the low-efficiency workers with young workers, Table 5 shows that this behavior is consistent with the dynamics of unemployment rates over the period 1983–2007 for which youth unemployment data are available. Volatility of youth and long-term unemployment is much higher in Euro area countries. The volatility of the youth unemployment rate is 200% higher than that of the aggregate unemployment rate in the EU-27 data, and only 32% higher in the U.S. data.

The middle left panel of Figure 5 shows that the log-increase in the unemployed share of low-efficiency workers peaks at 6%, or about a third of its increase in the EU case, limiting the relevance of the composition effect for unemployment volatility. Low-efficiency workers experience a pronounced fall in average hours relative to high-efficiency workers in both the EU and U.S. cases, a result consistent with the empirical evidence in Bils, Chang, and Kim (2009) and Hines, Hoynes, and Krueger (2001) that an important fraction of the fall in wage earnings for workers with below-average wages during a recession comes from a fall in hours worked.

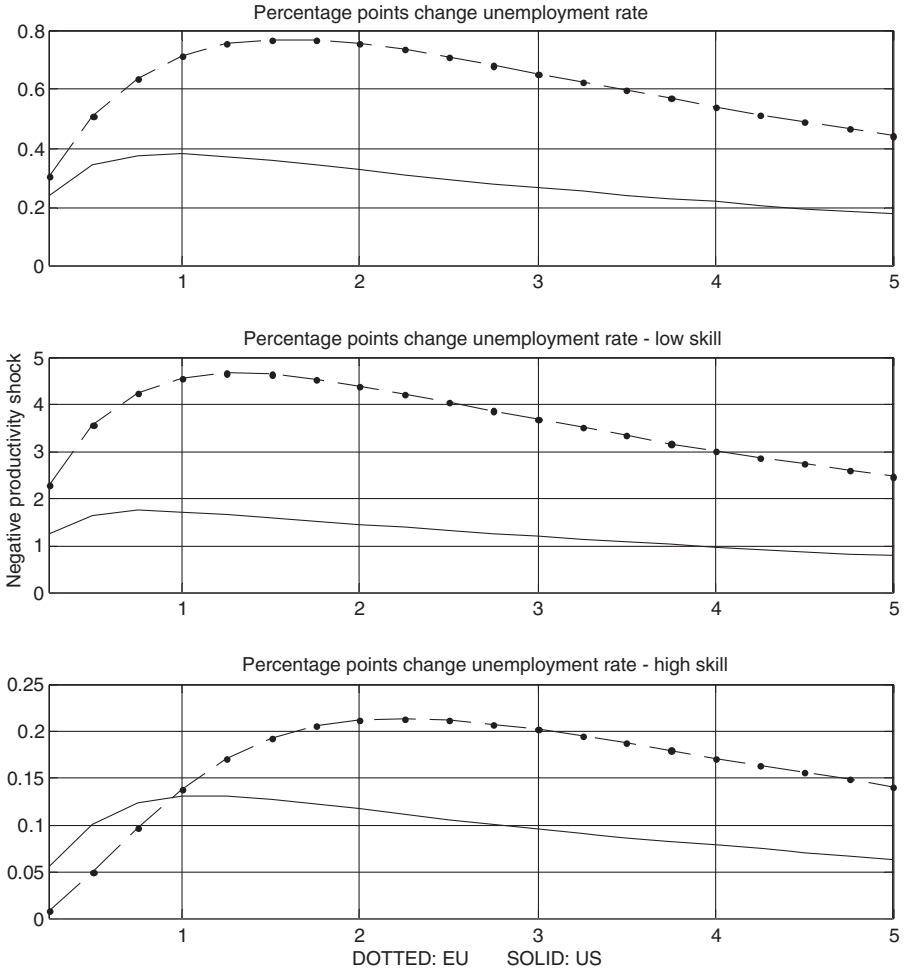

FIG 4. Impulse Response to a 1% Negative TFP shock z_t Under Price Stability for the Baseline Parameterization (EU) and a Parameterization Matching U.S. Data, Described in Table 4.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Change in unemployment rate for total, low- and high-efficiency population scaled in percentage points of the labor force L, L^l, L^h of each group. Horizontal axis in years.

2.4 The Impact of a Productivity Shock Biased against Low-Efficiency Workers

We next consider the impact of a fall in productivity that disproportionately affects low-efficiency workers. For this experiment we use the U.S. parameterization to show that even in an economy with large steady-state labor reallocation, a productivity shock biased against low-efficiency workers can substantially amplify unemployment volatility. The shock results in a large surge of low-efficiency workers into

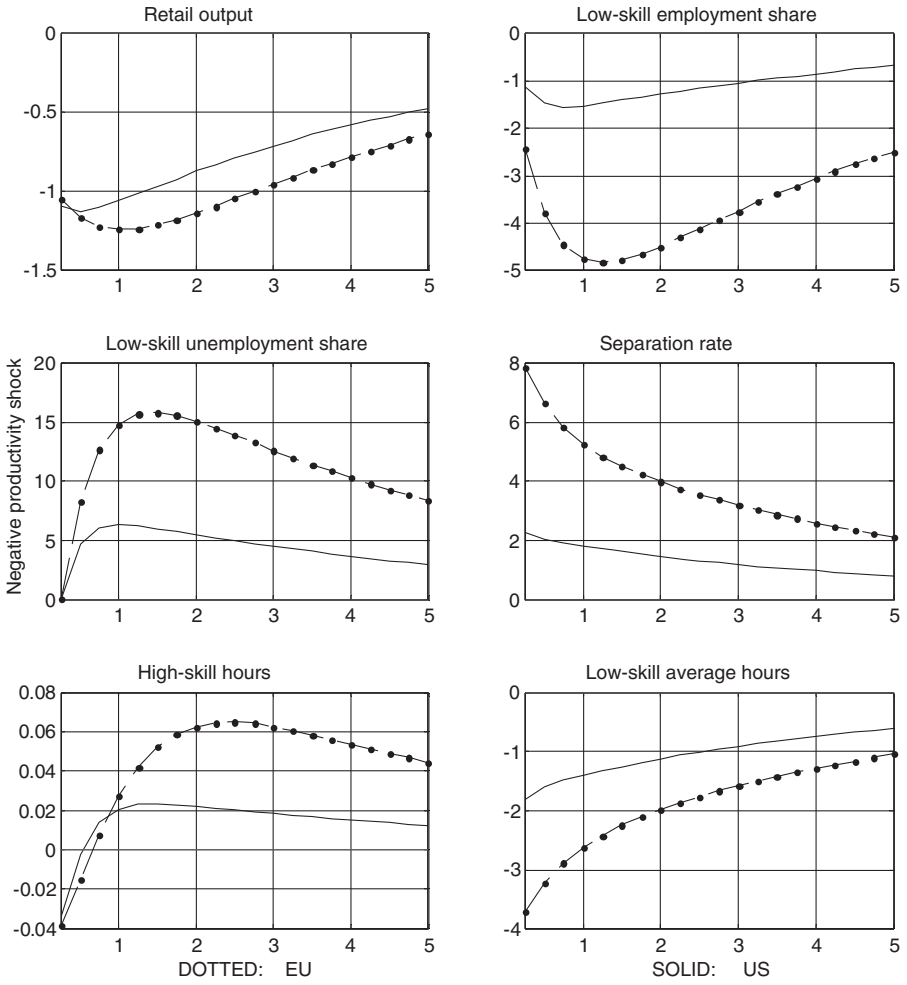

FIG 5. Impulse Response to a 1% Negative TFP Shock z_t Under Price Stability for the Baseline Parameterization (EU) and a Parameterization Matching U.S. Data, Described in Table 4.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Scaling in percent. Horizontal axis in years.

unemployment and a large increase in the low-efficiency share of unemployment, which then takes a long time to revert to its steady-state value.

We compare a 1% fall in aggregate TFP for the U.S. parameterization with a TFP shock affecting predominantly the low-efficiency labor force. We scale this asymmetric shock so that the initial decline in output is the same obtained in response to the aggregate TFP shock. This is achieved with a 0.7% decline in productivity of the high-efficiency workers, and a decline that is four times as large for the low-efficiency workers. While this may seem a large bias in the TFP shock, recall that the

TABLE 5

UNEMPLOYMENT RATE, 1983–2007

	Average	Standard deviation
Euro area		
Unemployment (% labor force)	10.11%	1.33
Unemployment: youth (% labor force ages 15–24)	22.16%	4.06
Unemployment: long term (% total unemployment)	48.74%	4.11
France		
Unemployment (% labor force)	9.98%	1.36
Unemployment: youth (% labor force ages 15–24)	22.32%	3.16
Unemployment: long term (% total unemployment)	40.47%	3.14
The U.S.		
Unemployment (% labor force)	5.84%	1.28
Unemployment: youth (% labor force ages 15–24)	12.03%	1.69
Unemployment: long term (% total unemployment)	9.25%	2.40

NOTE: Annual data.
SOURCE: World Development Indicators (2009).

share of l workers in the labor force is only 16%, so the large fall in TFP is affecting a small fraction of workers.

Even though the asymmetric shock generates a similar fall in output as the aggregate TFP shock on impact, the top left panel of Figure 6 shows that it generates a rise in unemployment over three times as large when compared to an aggregate shock. The unemployment increase is also greatly amplified for high-efficiency workers even though they experience a fall in TFP equal to only 70% of the fall in the case of an aggregate productivity shock. This amplification is due entirely to the impact of the productivity decline of low-efficiency workers on aggregate variables. While the difference in output is smaller across the aggregate and asymmetric shocks—since the bulk of employed workers belong to the high-efficiency group—the impact on unemployment is radically different.

Figure 7 illustrates the channels through which the large amplification in unemployment is obtained: the separation rate increases sharply, raising the share of low-efficiency unemployed in the jobless pool by 27% relative to the steady state. These workers, in turn, face a smaller chance of exiting unemployment, keeping the unemployment rate high for a prolonged period. Low-efficiency workers who remain employed also optimally lower their hours worked, although this fall in hours plays a small role in the fall in output, given the small share of low-efficiency workers in productive matches.

Finally, the middle left panel of Figure 7 plots the vacancy yield normalized by the number of unemployed workers. The distance between the two impulse responses is a measure of the shortfall in vacancy yield when the TFP shock is biased against one type of workers. The shortfall in the aggregate vacancy yield has been documented in the recent U.S. recession by Daly, Hobijn, and Valletta (2011). Heterogeneity in efficiency levels amplifies the fall in vacancy yield by a factor of 6.

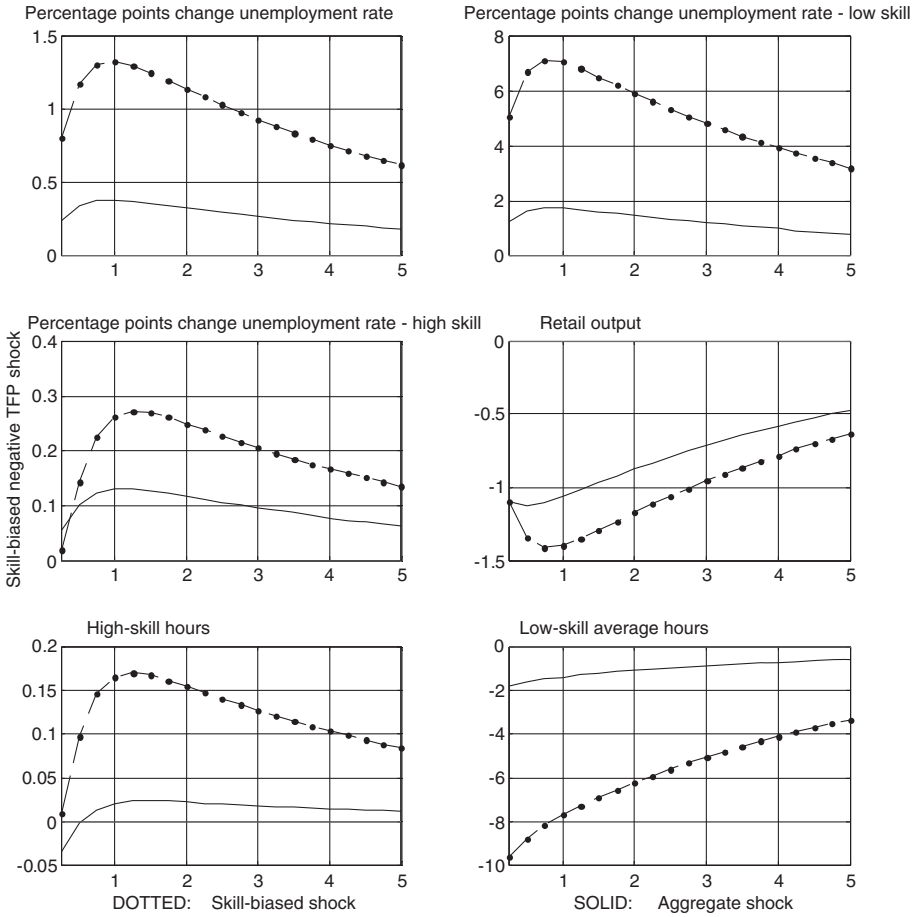

FIG 6. Impulse Response to a Negative TFP Shock z_t Biased Against Low-Efficiency Workers Under Price Stability for the Parameterization Matching U.S. Data, Described in Table 4.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. TFP innovation is equal to -0.7% for high-efficiency workers, and -2.8% for low-efficiency workers. For the case of an aggregate TFP shock, innovation is equal to -1%. Change in unemployment rate for total, low- and high-efficiency population scaled in percentage points of the labor force L, L^l, L^h of each group. Horizontal axis in years.

2.5 The Impact of Monetary Policy in an Economy with Heterogeneous Worker Productivity

The presence of nominal rigidities allows the monetary authority to affect the equilibrium dynamics in response to business cycle shocks. This section analyzes the impact of expansionary monetary policies that attempt to reduce the rise in unemployment of a recessionary productivity shock rather than simply to stabilize the price level. We compare for our baseline EU parameterization the flexible-price equilibrium delivered by a policy of price stability with the equilibrium conditional

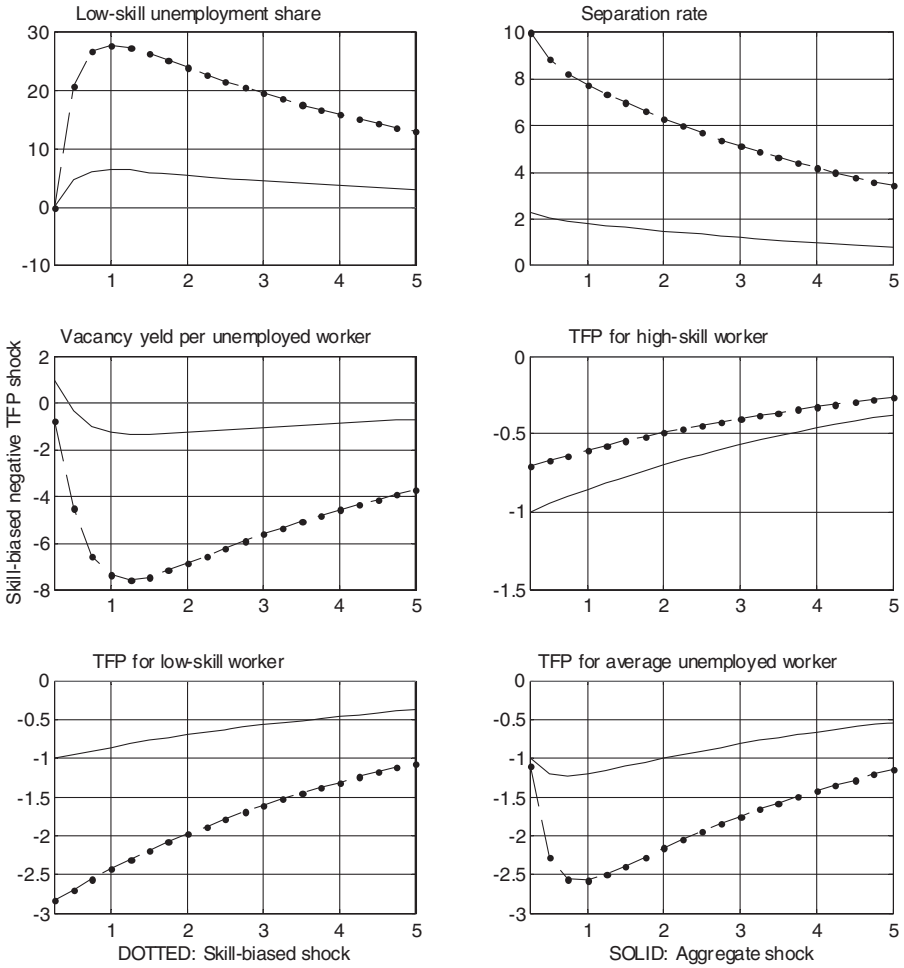

FIG 7. Impulse Response to a Negative TFP Shock z_t Biased against Low-Efficiency Workers Under Price Stability for the Parameterization Matching U.S. Data, Described in Table 4.

NOTES: AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. TFP innovation is equal to -0.7% for high-efficiency workers, and -2.8% for low-efficiency workers. For the case of an aggregate TFP shock, innovation is equal to -1%. Scaling in percent. Horizontal axis in years.

on a policy that lowers unemployment by approximately half at the peak of the recession. The expansionary policy is described by a simple Taylor-type instrument rule reacting to retail inflation and the unemployment rate:[13]

$$\ln(1 + i_t) = -\ln \beta + \omega_\pi \pi_t - \omega_u (U_t - \overline{U}). \tag{24}$$

13. Recall that $U_t \equiv 1 - N_t$.

We assume $\omega_\pi = 1.5$, $\omega_u = 0.1$; these values achieve a 50% reduction in unemployment in the face of a negative TFP shock while at the same time allowing policy to limit the volatility of inflation. Taylor-type rules incorporating a response to a measure of real economic activity have been considered by many authors. Orphanides and Williams (2002) examine the performance of a policy rule specified as in (24) where the interest rate reacts to deviations of the unemployment rate from a slow-moving estimate of the natural rate, rather than to a constant value, as in our approach.

Figures 8 and 9 show the impact of an expansionary policy in response to a fall in aggregate TFP. After one quarter, the expansionary policy is able to lower the severity of the recession. From the third quarter onward, the gain in output for the expansionary policy stabilizes around 0.3% of the steady-state value. Since the output response is very long lived, the cumulative gain in output is substantial under this policy. Part of the reduction in employment volatility with an expansionary monetary policy comes from the smaller size of the composition effect. Since the increase in the low-efficiency unemployment share in total unemployment is less than half as large, compared to the price stability policy, the fall in TFP for the average unemployed worker is proportionately reduced. In equilibrium, the expansionary policy leads to both a smaller rise in separations, and a smaller decline in job-finding probability. The reduction in unemployment is proportionally distributed across the efficiency groups. The expansionary policy also leads to a faster recovery, with unemployment peaking after four quarters.

The greater stability of output and employment is achieved at a large cost in terms of inflation, however, which jumps on impact by about 3.5%, and after 2 years is still 3% above steady state (see the upper right panel of Figure 8). The impact of monetary policy on unemployment can be illustrated through (4) and (5), describing the surplus of a match for a high- and low-efficiency worker. An expansionary monetary policy leads to temporarily smaller markup μ_t since retail prices are sticky. Markups affect the surplus value symmetrically but in the opposite direction of technology shocks. From the point of view of the intermediate firm–worker match, the value of the match increases in terms of retail goods as markups fall (both in terms of consumption value for the worker, and in terms of revenue for the intermediate goods-producing firm). A lower markup translates into a temporarily higher marginal cost for the retail firms, leading to pressure for price increases and to inflation.[14] As markups fall, the increase in the separation rate is reduced relative to the case under price stability. This in turn leads to a smaller decline in the average efficiency of the unemployed workers, restraining the amplification on TFP for the unemployed pool generated by the composition effect, and the corresponding fall in job-finding probability.

Figures 10 and 11 compare the two monetary policies for the case of a TFP shock biased against low-efficiency workers. To allow comparability with the asymmetric

14. Ravenna and Walsh (2012) discuss in depth the interpretation of monetary policy as a tax/subsidy to intermediate firms' revenues.

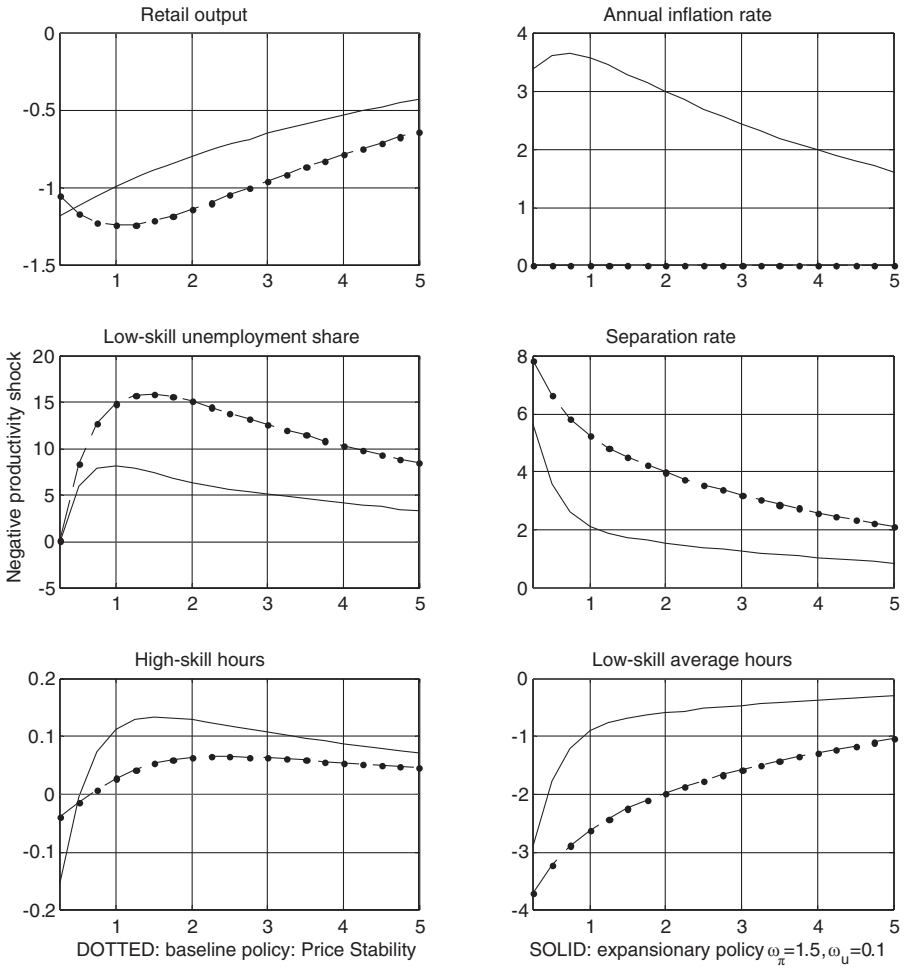

FIG 8. Impulse Response to a 1% Negative TFP Shock z_t Under the Baseline Parameterization (EU) for Alternative Policies.

NOTES: The expansionary policy rule is described in equation (24), and assumes $\omega_\pi = 1.5,\ \omega_u = 0.1$. AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Horizontal axis in years.

shock impulse response shown in Figures 6 and 7, we adopt the U.S. parameterization. While we do not assess the welfare impact of alternative monetary policies, it is clear that the inflation–unemployment trade-off for a shock biased against low-efficiency workers is much worse than for an aggregate TFP shock, making monetary policy a very ineffective tool in responding to the recessionary shock. To generate an approximate 50% fall in the impact of the shock on unemployment, the annual inflation rate has to increase by nearly 11 percentage points above the steady state. After eight quarters, the inflation rate is still 8% above steady state. The increase in

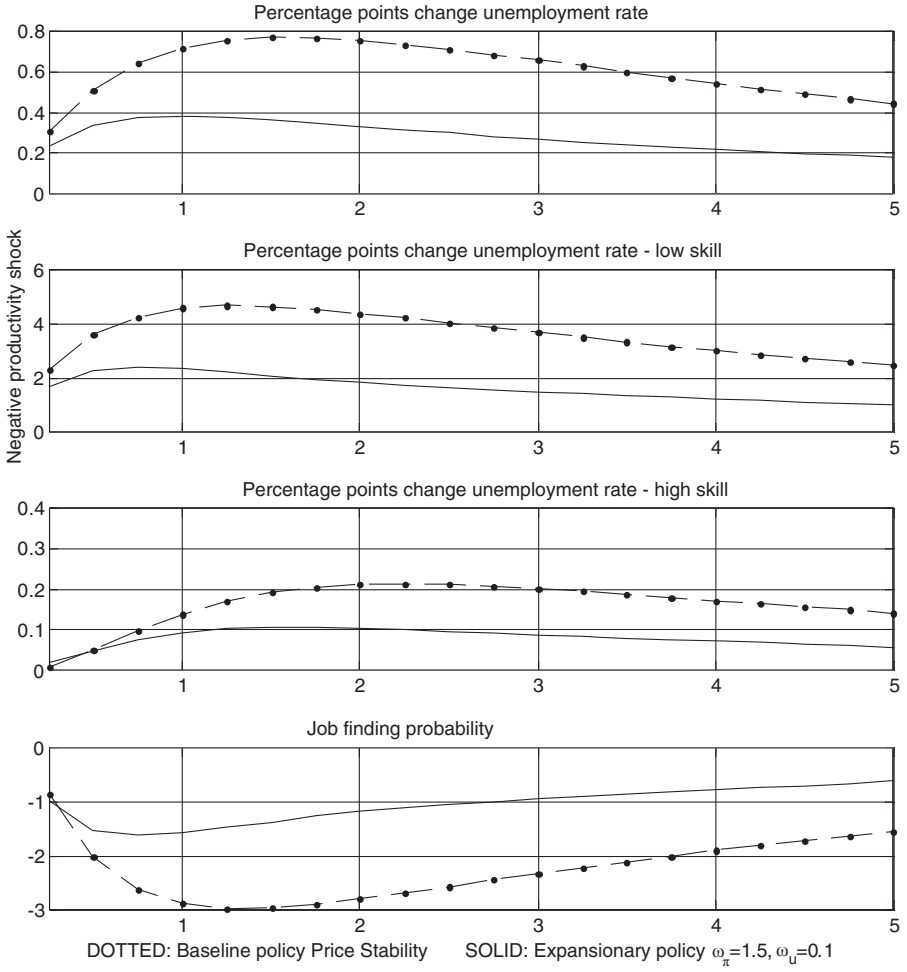

FIG 9. Impulse Response to a 1% Negative TFP Shock z_t Under the Baseline Parameterization (EU) for Alternative Policies.

NOTES: The expansionary policy rule is described in equation (24), and assumes $\omega_\pi = 1.5,\ \omega_u = 0.1$. AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. Horizontal axis in years.

unemployment falls disproportionately on the low-efficiency group. Since this is the group for which TFP falls by a larger amount, increasing the surplus of low-efficiency workers requires a much larger fall in the aggregate markup compared to the case of an aggregate TFP shock. In turn, the policy is more effective in reducing the unemployment rate among high-efficiency workers, as shown in Figure 11. At peak, the expansionary policy can reduce the TFP fall for the average unemployed worker only by half a percentage point, from 2.5% to 2%.

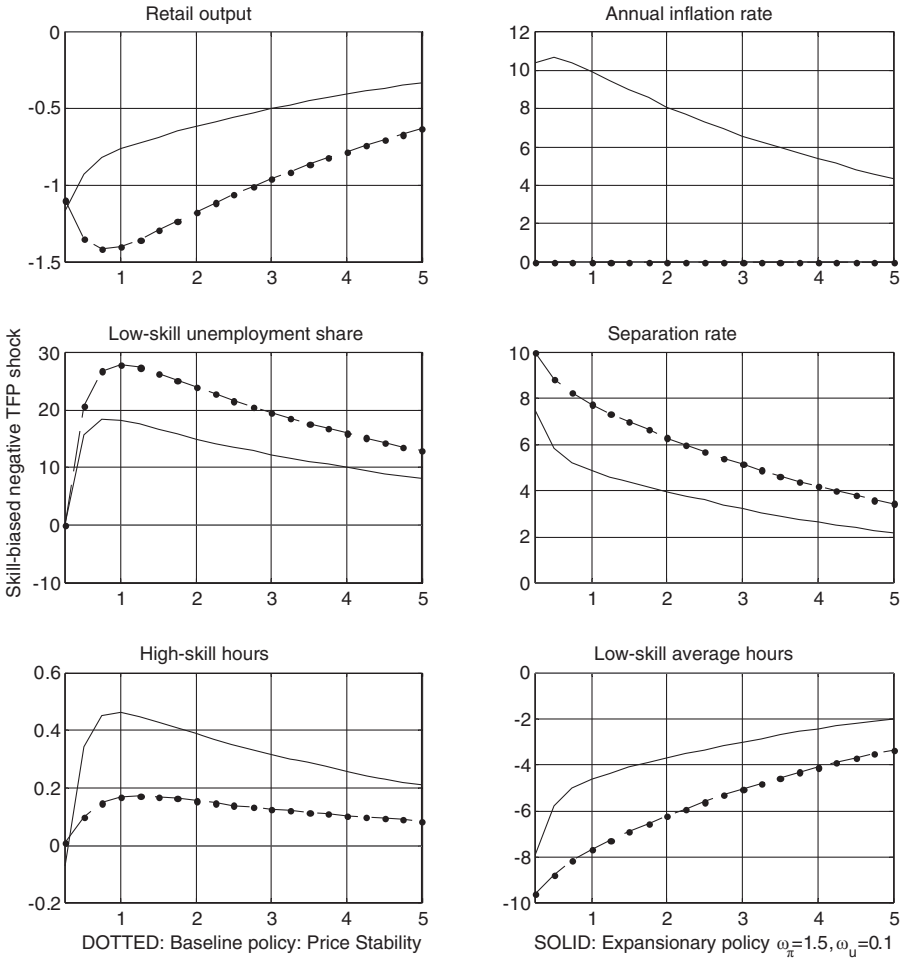

FIG 10. Impulse Response to a Negative TFP Shock z_t Biased against Low-Efficiency Workers Under the Parameterization Matching U.S. Data, Described in Table 4, for Alternative Policy Rules.

NOTES: The policy rule is described in equation (24), and assumes $\omega_\pi = 1.5,\ \omega_u = 0.1$. AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. TFP innovation is equal to -0.7% for high-efficiency workers, and -2.8% for low-efficiency workers. Horizontal axis in years.

In the case of a negative asymmetric shock, the policy that responds directly to the rise in unemployment is more effective in reducing the decline in output than it is in reducing the decline in the unemployment rate. This is the consequence of an increase in the number of hours optimally supplied by high-efficiency workers, which represent the bulk of the employed workforce.[15]

15. The differential impact on employment and output can be explained by considering the impact of a change in markups on the surplus value (equations (4) and (5)) and hours worked (equations (10) and (11)), showing that the value of the surplus and of hours do not have to move proportionally.

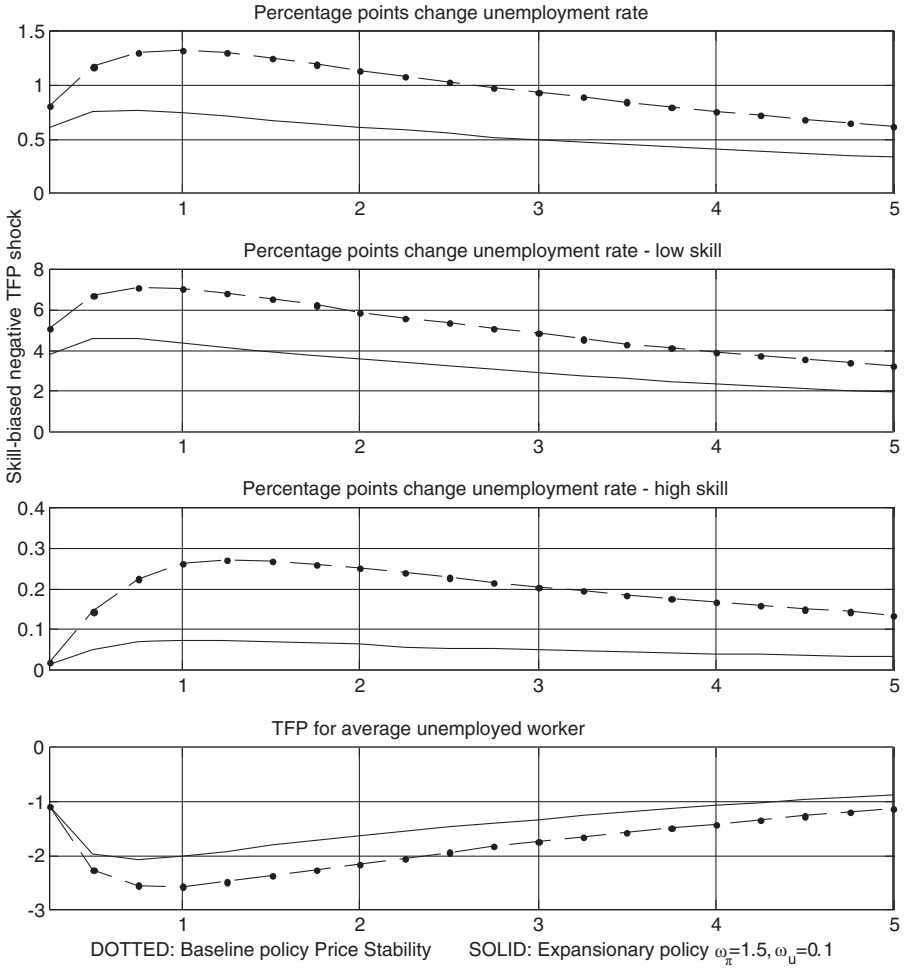

FIG 11. Impulse Response to a Negative TFP Shock z_t Biased against Low-Efficiency Workers Under the Parameterization Matching U.S. Data, Described in Table 4, for Alternative Policy Rules.

NOTES: The policy rule is described in equation (24), and assumes $\omega_\pi = 1.5$, $\omega_u = 0.1$. AR(1) coefficient of TFP shock $\rho_{z_t} = 0.95$. TFP innovation is equal to -0.7% for high-efficiency workers, and -2.8% for low-efficiency workers. Horizontal axis in years.

3. EMPIRICAL EVIDENCE: HETEROGENEITY IN WORKER PRODUCTIVITY AND AGGREGATE LABOR MARKET DYNAMICS

The hypothesis that changing heterogeneity in the pool of unemployed may drive the correlation between aggregate labor market variables and the business cycle has a long history. Darby, Haltiwanger, and Plant (1985) advance the heterogeneity hypothesis to explain the strong countercyclicality of average unemployment duration: if the

composition of job losers changes systematically over the business cycle, and groups that experience longer durations enter unemployment in proportionally greater numbers during a recession, the average spell will be countercyclical even if individual spells are acyclical.

Most empirical studies of the heterogeneity hypothesis have relied on observable heterogeneity. Baker (1992) finds no support for the heterogeneity hypothesis in unemployment duration when selecting groups by demographics or reason for joblessness, and similar results are obtained for unemployment exit rates by Abbring, van den Berg, and van Ours (2002) and van den Berg and van der Klaauw (2001) using French data. One observable characteristic that has received much attention in the literature as a potential driver of time-varying heterogeneity in the unemployment pool is the reason for joblessness. Davis, Haltiwanger, and Schuh (1996) provide evidence supporting the hypothesis that a disproportionate part of unemployment inflows during a recession consists of laid-off workers, and stress the countercyclical behavior of layoffs, as opposed to the procyclical behavior of quits (which reflect in large part job-to-job transitions). The recent literature has stressed that the separation rate is rather acyclical in recent U.S. business cycles, but the data show that increased inflows into unemployment during a recession can be traced to a shift in separations toward layoffs (Elsby, Hobijn, and Sahin 2008, Ramey and Fujita 2009). Our model is consistent with this evidence, since the share of endogenous separation is procyclical, leading to low-efficiency workers being overrepresented in the group flowing into unemployment during a recession. Davis (2005) cites several studies finding that layoffs are associated with greater unemployment incidence and longer unemployment spells than quits, and workers experiencing layoffs also experience a large and persistent decline in earnings.

Using U.S. Current Population Survey (CPS) data from 1976 to 2007, Shimer (2012) reports that, while the change in the share of laid-off workers is correlated with the business cycle, it explains a small portion of the overall variation in the job-finding probability. Similarly, the data in Elsby, Hobijn, and Sahin (2010) show that the bulk of the large differences in the level of unemployment across demographic subgroups are driven by differences in each group's inflow rate. Outflow rates from unemployment are remarkably more similar than inflow rates by age, education, ethnicity. We provide a theoretical framework that is partly consistent with this evidence by allowing both high- and low-efficiency workers to compete in the same labor market. Thus, while we assume inflows into unemployment increases only for low-efficiency workers, the outflow rate endogenously falls for all workers—though proportionally more for low-efficiency workers—as the composition effect reduces the incentive of firms to post vacancies.

Barnichon and Figura (2011) provide evidence in support of the heterogeneity hypothesis. They examine the role of heterogeneity in explaining changes in matching efficiency for a matching function estimated using CPS data for the 1976–2009 sample. Their estimates support the finding that most of the shifts in the matching function up to 2006, and half of the decline in matching efficiency over the 2007–09 period, are due to changes in the composition of the pool of unemployed.

The model we propose relies on unobserved heterogeneity, as workers with heterogeneous productivity cannot be sorted according to observable characteristics before being interviewed by a firm. Thus, empirical studies that examine the heterogeneity hypothesis relying on demographic data for age or education, or looking at sectorial data, do not provide direct support for our assumptions. Unobserved heterogeneity and its relation with the behavior of aggregate labor market variables over the business cycle has been considered by some authors. Many labor economists have documented that most of the wage differentials across workers cannot be explained by observable characteristics.[16] Education and experience are often badly measured and do not fully capture the effectiveness of a worker. Pries (2008) observes that efficiency heterogeneity is hard to measure, both because a worker's productivity is only partially accounted for by observable characteristics and because workers can differ in the value of their outside option relative to employment. Abbring, van der Berg, and van Ours (2002) find using French data that unemployment duration dependence over the first five quarters is explained by unobserved heterogeneity.

Mueller (2011) provides some direct evidence related to our assumption of unobserved heterogeneity. Using CPS data, he shows that separation and job-finding rates are more cyclical for high-residual wage workers as opposed to low-residual wage ones. If we attribute the above-median wage residual to a higher efficiency level, the evidence points toward a reverse impact of heterogeneity than the one assumed in our model. However, our model can match his evidence on the procyclicality of wages for workers flowing into the pool of unemployed, since it implies that in the beginning of a recession the productivity of low-efficiency workers entering unemployment is higher than average, and the average wage for unemployment entrants does not need to fall. Bils, Chang, and Kim (2009) find results opposite to Mueller using U.S. Survey of Income and Program Participation (SIPP) data. They conclude that low-wage, low-hours workers (which they identify with workers having a low comparative advantage on the labor market in comparison to nonmarket activities) have separation and job-finding rates substantially more sensitive to the business cycle than high-wage, high-hours workers.

4. CONCLUSIONS AND EXTENSIONS

We have developed a parsimonious model of worker heterogeneity that incorporates endogenous separations. Heterogeneity causes the composition of the pool of unemployed workers to vary over the business cycle in ways that cannot occur in standard models with homogenous labor. A negative productivity shock reduces output and employment, but it also lowers the average quality of the unemployed,

16. See Mortensen (2003). A substantial literature examines the behavior of wages over the business cycle, and has considered the heterogeneity hypothesis (see, e.g., Solon, Barski, and Parker 1994).

as low-efficiency workers experience a greater inflow into unemployment. This composition effect reduces the incentive for firms to post vacancies, as they are less likely to find a worker who is sufficiently productive to generate a positive surplus if hired. As a consequence, the exit rate from unemployment falls for all workers relative to a model with homogeneous labor.

As den Haan, Ramey, and Watson (2000) show, endogenous separation can contribute to both the amplitude of employment responses to productivity shocks and the persistence generated by such shocks. We find that these effects are further strengthened by compositional affects that arise with heterogeneous workers. Despite the introduction of only two worker types, the model generates a rich set of implications for unemployment inflows and outflows. Heterogeneity in worker efficiency amplifies unemployment fluctuations in economies with small gross labor flows and lowers the vacancy yield during recessions. It additionally results in a slow buildup of unemployment, peaking seven quarters after the beginning of the recession, and a slow recovery. Even in economies with large average worker turnover rates, a productivity shock that disproportionally affects workers of low efficiency from the point of view of prospective employer, results in a very large increase in the relative volatility of employment to output. All of these results are obtained with a low-efficiency share of the labor force of the order of 15%, and with relatively small changes in the efficiency composition of unemployment.

The model provides a platform on which to investigate the role of labor market dynamics in affecting the transmission of monetary policy, and the effects of macroeconomic fluctuations on unemployment flows in different countries or global regions characterized by different labor market structures.

We believe models with workers heterogeneity raise very important questions for monetary policy. We considered the impact on unemployment stabilization of two simple rules for monetary policy. However, as discussed in Ravenna and Walsh (2011), in search and matching models unemployment stabilization is not an optimal policy. As is well known, a form of congestion externality is present in search and matching models; a firm that posts a vacancy reduces the probability other firms are able to fill their vacancies. With worker heterogeneity and endogenous separations, an additional externality arises. When a firm fails to retain a low-efficiency worker, the average productivity of the pool of job seekers is lowered, thus making it less likely that a firm with a vacancy will make a hire. And as firms hire high-efficiency workers, they increase the probability that other firms will end up with a low-efficiency worker. While in our model workers differ only in the amount of efficiency units they can provide, additional externalities arise in models with sorting, where both jobs and workers are heterogeneous and there exist complementarities between jobs' and workers' characteristics. In these models, efficiency no longer simply requires that the appropriate number of workers is employed but also requires that workers are employed at the right jobs (Blázquez and Jansen 2008). The impact of these externalities on optimal monetary policy is left open for future research.

LITERATURE CITED

Abbring, Jaap, Gerard van den Berg, and Jan van Ours. (2002) "The Anatomy of Unemployment Dynamics." *European Economic Review*, 46, 1785–1824.

Baker, Michael. (1992) "Unemployment Duration: Compositional Effects and Cyclical Variability." *American Economic Review*, 82, 315–21.

Barnichon, Regis, and Andrew, Figura. (2011) "What Drives Matching Efficiency? A Tale of Composition and Dispersion." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series 2011–10.

Bils, Mark, Yongsung Chang, and Sun-Bin Kim. (2009) "Comparative Advantage and Unemployment." NBER Working Paper No. 15030.

Blanchard, Olivier, and Jordi Galí. (2007) "Real Wage Rigidity and the New Keynesian Model." *Journal of Money, Credit and Banking*, 39, 18–45.

Blanchard, Olivier, and Jordi Galí. (2010) "A New Keynesian Model with Unemployment." *American Economic Journal: Macroeconomics,* 2, 1–30.

Blázquez, Maite, and Marcel Jansen. (2008) "Search, Mismatch and Unemployment." *European Economic Review*, 52, 498–526.

Clark, Kim B., and Lawrence H. Summers. (1981) "The Demographic Composition of Cyclical Employment Variations." *Journal of Human Resources*, 16, 61–79.

Daly, M., Bart Hobijn, and Robert Valletta. (2011) "The Recent Evolution of the Natural Rate of Unemployment." FRB San Francisco Working Paper No. 2011–05.

Darby, Michael R., John C. Haltiwanger, and Mark Plant. (1985) "Unemployment Rate Dynamics and Persistent Unemployment under Rational Expectations." *American Economic Review*, 75, 614–37.

Davis, Steven T. (2005) "Job Loss, Job Finding, and Unemployment in the U.S. Economy over the Past Fifty Years. Comment." *NBER Macroeconomics Annual*, 20, 139–57.

Davis, Steven T., John C. Haltiwanger, and Scott Schuh. (1996) *Job Creation and Job Destruction*, Cambridge, MA: MIT Press.

den Haan, Wouter J., Garey Ramey, and Joel Watson. (2000) "Job Destruction and Propagation of Shocks." *American Economic Review*, 90, 482–98.

Elsby, Michael, Bart Hobijin, and Aysegul Sahin. (2008) "Unemployment Dynamics in the OECD." NBER Working Paper No. 14617.

Elsby, Michael, Bart Hobijin, and Aysegul Sahin. (2010) *Brookings Papers on Economic Activity*, Spring, 1–48.

Galí, Jordi. (2011) *Unemployment Fluctuations and Stabilization Policies: A New Keynesian Perspective.* Cambridge, MA: MIT Press.

Gertler, Mark, and Antonella Trigari. (2009) "Unemployment Fluctuations with Staggered Nash Wage Bargaining." *Journal of Political Economy*, 117, 38–86.

Gertler, Mark, Luca Sala, and Antonella Trigari. (2008) "An Estimated Monetary DSGE Model with Unemployment and Staggered Nominal Wage Bargaining." *Journal of Money, Credit and Banking,* 40, 1713–64.

Guerrieri, Veronica. (2007) "Heterogeneity and Unemployment Volatility." *Scandinavian Journal of Economics*, 109, 667–93.

Hines, James R., Hilary Hoynes, and Alan B. Krueger. (2001) "Another Look at Whether a Rising Tide Lifts All Boats." In *The Roaring Nineties: Can Full Employment Be Sustained?* edited by Alan B. Krueger and Robert Solow, pp. 329–51. New York: Russell Sage Foundation.

Hornstein, Andreas, Per Krusell, and Giovanni L. Violante. (2007) "Frictional Wage Dispersion in Search Models: A Quantitative Assessment." Federal Reserve Bank of Richmond Working Paper No. 06–10.

Jung, Philip, and Moritz Kuhn. (2011) "Labor Market Rigidity and Business Cycle Volatility." Mimeo, University of Mannheim.

Krause, Michael, and Thomas Lubik. (2010) "On-the-Job Search and the Cyclical Dynamics of the Labor Market." Federal Reserve Bank of Richmond Working Paper No. 10–12.

Learning and Skills Council. (2008) *National Employers Skills Survey 2007: Key Findings*. Coventry, UK: Learning and Skills Council National Office.

Manning, Alan. (2011) "Imperfect Competition and the Labour Market." In *Handbook of Labor Economics,* edited by Orley Ashenfelter and David Card, pp. 205–255. London: Elsevier.

Mortensen, Dale T. (2003) *Wage Dispersion: Why Are Similar People Paid Differently.* Cambridge, MA: MIT Press.

Mueller, Andreas. (2011) "Separations, Sorting and Cyclical Unemployment." Mimeo, Columbia University.

Nagypal, Eva. (2007) "Learning-by-Doing Versus Learning about Match Quality: Can We Tell Them Apart?" *Review of Economic Studies* , 74, 537–66.

Nagypal, Eva, and Dale T. Mortensen. (2007) "Labor-Market Volatility in Matching Models with Endogenous Separations." *Scandinavian Journal of Economics*, 109, 645–65.

Orphanides, Athanasios, and John C. Williams. (2002) "Robust Monetary Policy Rules with Unknown Natural Rates." *Brookings Papers on Economic Activity*, Spring, 63–118.

Petrongolo, Barbara, and Christopher Pissarides. (2001) "Looking into the Black Box: A Survey of the Matching Function." *Journal of Economic Literature*, 39, 390–431.

Pries, Michael. (2008) "Worker Heterogeneity and Labor Market Volatility in Matching Models." *Review of Economic Dynamics*, 11, 644–87.

Pries, Michael, and Richard Rogerson. (2005) "Hiring Policies, Labor Market Institutions, and Labor Market Flows." *Journal of Political Economy,* 113, 811–39.

Ramey, Garey, and Shigeru Fujita. (2009) "The Cyclicality of Separation and Job Finding Rates." *International Economic Review*, 50, 415–30.

Ravenna, Federico, and Carl E. Walsh. (2008) "Vacancies, Unemployment, and the Phillips Curve." *European Economic Review*, 52, 1120–45.

Ravenna, Federico, and Carl E. Walsh. (2011) "Welfare-Based Optimal Monetary Policy with Unemployment and Sticky Prices: A Linear-Quadratic Framework." *American Economic Journal: Macroeconomics*, 3, 320–43.

Ravenna, Federico, and Carl E. Walsh. (2012) "The Welfare Consequences of Monetary Policy and the Role of the Labor Market: A Tax Interpretation." *Journal of Monetary Economics*, 59, 180–95.

Saratoga Institute. (2000) *Saratoga Institute Human Resource Financial Report*. Saratoga: Saratoga Institute.

Shimer, Robert. (2005) "The Cyclical Behavior of Equilibrium Unemployment and Vacancies." *American Economic Review*, 95, 25–49.

Shimer, Robert . (2012) "Reassessing the Ins and Outs of Unemployment." *Review of Economic Dynamics*, 15, 127–48.

Solon, Gary, Robert Barski, and Jonathan A. Parker. (1994) "Measuring the Cyclicality of Real Wages: How Important Is Composition Bias?" *Quarterly Journal of Economics,* 109, 1–28.

Tasci, Murat. (2007) "On the Job Search and Labor Market Reallocation." Federal Reserve Bank of Cleveland Working Paper No. 7–25.

van den Berg, Gerard J., and Bas van der Klaauw. (2001) "Combining Micro and Macro Unemployment Duration Data." *Journal of Econometrics,* 102, 271–309.

Villena-Roldan, Benjamin. (2008) "Aggregate Implications of Employer Search and Recruiting Selection." Mimeo, University of Rochester.

Walsh, Carl E. (2003) "Labor Market Search and Monetary Shocks." In *Elements of Dynamic Macroeconomic Analysis*, edited by Sumru Altuĝ, Jagit S. Chadha, and Charles Nolan, pp. 451–86. Cambridge, UK: Cambridge University Press.

Walsh, Carl E. (2005) "Labor Market Search, Sticky Prices, and Interest Rate Policies." *Review of Economic Dynamics*, 8, 829–49.