DISCUSSION PAPER SERIES

DP13628
(v. 5)

Bad Jobs and Low Inflation

Leonardo Melosi and Renato Faccini

MONETARY ECONOMICS AND FLUCTUATIONS

CEPR

ISSN 0265-8003

Bad Jobs and Low Inflation

Leonardo Melosi and Renato Faccini

Discussion Paper DP13628
First Published 27 March 2019
This Revision 10 February 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Monetary Economics and Fluctuations

Bad Jobs and Low Inflation

Abstract

The low rate of inflation observed in the U.S. over the entire past decade is hard to reconcile with traditional measures of labor market slack. We show that an alternative notion of slack that encompasses workers' propensity to search on the job explains this missing inflation. We derive this novel concept of slack from a model in which a drop in the on-the-job search rate lowers the intensity of interfirm wage competition to retain or hire workers. The on-the-job search rate can be measured directly from aggregate labor-market flows and is countercyclical. Its recent drop is corroborated by micro data.

JEL Classification: E31, E37, C32

Keywords: Missing inflation, on-the-job search, employment-to-employment rate, labor market slack, Phillips curve, Cyclical Misallocation, Micro data, Heterogeneous Agents

Leonardo Melosi - lmelosi@frbchi.org
Federal Reserve Bank of Chicago and CEPR

Renato Faccini - r.faccini@qmul.ac.uk
Danmarks Nationalbank

Acknowledgements
Correspondence to: rmmf@nationalbanken.dk and lmelosi@frbchi.org. We thank Gadi Barlevy, Robert Barsky, Marco Bassetto, Lawrence Christiano, Martin Eichenbaum, Jason Faberman, Filippo Ferroni, Jonas Fisher, Nir Jaimovich, Michael Krause, Giuseppe Moscarini, Fabien Postel-Vinay, Sergio Rebelo, and seminar participants at the Society for Economic Dynamics, Boston Fed, Chicago Fed, European University Institute, LSE, Northwestern University, Bank of England, University of Warwick, and the EES conference in Stockholm on New Developments in the Macroeconomics of Labor Markets for their comments and suggestions. We also thank Jason Faberman for sharing the series of quit rates for the 1990s and May Tysinger for providing excellent research assistance. The views in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of Chicago, Danmarks Nationalbank, or any person associated with the Federal Reserve System or the European System of Central Banks.

Bad Jobs and Low Inflation*

Renato Faccini
Danmarks Nationalbank
Centre for Macroeconomics (LSE)

Leonardo Melosi
FRB Chicago
European University Institute
CEPR

February 9, 2021

Abstract

The low rate of inflation observed in the U.S. over the entire past decade is hard to reconcile with traditional measures of labor market slack. We show that an alternative notion of slack that encompasses workers' propensity to search on the job explains this missing inflation. We derive this novel concept of slack from a model in which a drop in the on-the-job search rate lowers the intensity of interfirm wage competition to retain or hire workers. The on-the-job search rate can be measured directly from aggregate labor-market flows and is countercyclical. Its recent drop is corroborated by micro data.

Keywords: Missing inflation, on-the-job search, employment-to-employment rate, labor market slack, Phillips curve, cyclical misallocation, micro data, heterogeneous agents.

JEL codes: E31, E37, C32

*Correspondence to: rmmf@nationalbanken.dk and lmelosi@frbchi.org. We thank Gadi Barlevy, Robert Barsky, Marco Bassetto, Lawrence Christiano, Martin Eichenbaum, Jason Faberman, Filippo Ferroni, Jonas Fisher, Nir Jaimovich, Michael Krause, Giuseppe Moscarini, Fabien Postel-Vinay, Sergio Rebelo, and seminar participants at the Society for Economic Dynamics, Boston Fed, Chicago Fed, European University Institute, LSE, Northwestern University, Bank of England, University of Warwick, and the EES conference in Stockholm on New Developments in the Macroeconomics of Labor Markets for their comments and suggestions. We also thank Jason Faberman for sharing the series of quit rates for the 1990s and May Tysinger for providing excellent research assistance. The views in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of Chicago, Danmarks Nationalbank, or any person associated with the Federal Reserve System or the European System of Central Banks.

"Our framework for understanding inflation dynamics could be misspecified in some fundamental way, perhaps because our econometric models overlook some factor that will restrain inflation in coming years despite solid labor market conditions."

Janet Yellen, Federal Reserve Chair, at the 59th Annual Meeting of the National Association for Business Economics in Cleveland, OH, on September 26, 2017

1 Introduction

Workhorse models used to study inflation attribute a key role to the labor market. When the labor market is tight, wage pressures and marginal costs increase, resulting in growing inflation; when the labor market is slack, wages and marginal costs fall and inflation decreases. This prediction is not borne out by the recent U.S. macroeconomic developments when the rate of unemployment is taken as a proxy for labor market slack, following a conventional approach dating back to Phillips (1958). As shown in Figure 1, from March 2017 through the end of 2019 the unemployment rate has stayed consistently below its average level measured over the last twelve months of the previous expansion, and by September 2019 it had reached its 50-year low at 3.5%. At the same time, core inflation according to the Price Index for Personal Consumption Expenditures (PCE) remained persistently below its long-term expectations.[1]

We first show that traditional measures of labor market slack fail to explain this missing inflation. We then introduce a more comprehensive measure of labor market slack, which depends on the unemployment rate as well as the employment-to-employment (EE) flow rate. As shown in Figure 1, the EE rate has recovered sluggishly in the previous decade, remaining below its pre-Great Recession average. One factor behind this low churn rate is the decline in the propensity of employed workers to search on the job, which we show to be supported by the micro data. We build a model to show that the low propensity to search on the job and the ensuing rise in low-productivity matches can explain the missing inflation in the last decade.

In the model, the productivity of jobs is match-specific and can be either high or low. All unemployed workers and a time-varying fraction of the employed search for a job. Firms have to compete to attract or retain workers who search on the job by bidding up their wage offers. As a result, these job seekers are more expensive to hire than the unemployed. A lower rate of on-the-job search reduces the incidence of wage competition between firms, leading to a decline in the expected labor costs and lower inflationary pressures. Intuitively, if firms expect their employees to be less willing to search and quit for another job, they will also anticipate less

[1] In the post-war period, the U.S. economy experienced low rates of unemployment and inflation in other periods, for example in the 1960s and in the 1990s. However, these episodes occurred in connection with high labor productivity growth, which in New Keynesian (NK) models lowers real marginal costs and hence dampens inflationary pressures. What made the latest expansion particularly puzzling was that inflation remained low while labor productivity growth also slowed down (Fernald 2016).

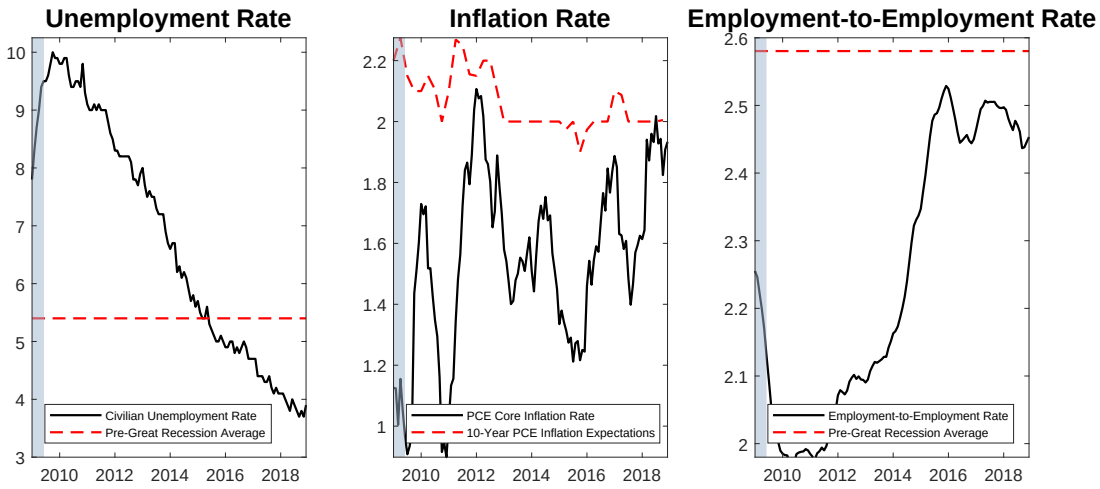

Figure 1: Labor market and inflation dynamics during the post-Great Recession recovery. The left panel: the civilian unemployment rate (solid line) and its average computed over a period spanning July 1990 through December 2007 (red dashed line). Source: Bureau of Labor Statistics (BLS). The central panel: the Price Index for Personal Consumption Expenditures (PCE) Excluding Food and Energy (black line). Average of monthly figures annualized in percentage. Source: Bureau of Economic Analysis (BEA). The red dashed line denotes the expectations about the PCE inflation rate over the next ten years. Source: Survey of Professional Forecasters. The right panel: employment-to-employment flow rate (black line) and its average computed over a period spanning July 1990 through December 2007 (red dashed line). Three-month centered moving average. Source: Current Population Survey (CPS) extended back to January 1990 by splicing it with the quit rate measured by Davis, Faberman, and Haltiwanger (2012) and corrected by Fujita, Moscarini, and Postel-Vinay (2019). The shaded areas denote the NBER recessions.

frequent pay raise requests to match outside offers and hence less pressure on payroll costs.

We first show that the on-the-job search rate in the model is implied by the unemployment rate and the EE flow rate and hence can be measured using aggregate labor market flows. A low and stagnant fraction of workers who switch jobs, combined with a tight labor market in which finding a job becomes easier over time, is interpreted by our model as a fall in the on-the-job search rate. To validate this prediction, we compute the on-the-job search rate at the micro level using the *Survey of Consumer Expectations* (SCE) administered by the Federal Reserve Bank of New York. In the Survey, the on-the-job search rate has fallen from 2014 through 2017 in a way that is remarkably similar to our measure based on the aggregate labor market flows. While the time period covered by the survey is quite short, it contains information on the search behavior of the employed over those years that are critical for our model to explain the missing inflation.

We derive a model-consistent concept of labor market slack, which can be measured using the observed series of the unemployment rate and the EE rate. Labor market slack hinges on the intensity of interfirm wage competition, which is shown to depend on (*i*) the unemployment rate, (*ii*) the degree of cyclical labor misallocation (i.e., the incidence of low-productivity jobs), and (*iii*) the on-the-job search rate. An increase in the rate of unemployment or a fall in the fraction of workers who are searching on the job increase slack in the model because they raise the firms' chances to fill their vacancies with unemployed workers, who are cheaper to hire, as they are unable to prompt wage competition between employers. The more inefficient the

allocation of labor, the more likely it is for firms to meet workers employed in low-productivity (bad) matches. Because enticing a worker away from a bad match is cheaper on average than poaching a worker from a good match, labor misallocation lowers the intensity of wage competition and raises labor market slack.

We then take the model to the unemployment rate and the EE flow rate observed in the data and recover the two shocks that buffet the model economy: a shock to the on-the-job search rate and a demand shock. The demand shock serves the sole purpose of generating the fluctuations in the unemployment rate observed in the data. Given the unemployment rate, the EE rate allows us to pin down the shocks to the on-the-job search rate, as described earlier. We use the time series of the two shocks to simulate inflation and labor costs from the model. We find that the model does not predict inflationary pressures during the most recent expansion, and this result is driven by the decline in the rate of on-the-job search, which has kept wage competition at low levels. In addition, labor-cost dynamics in the model closely correlate with the growth rate of average hourly earnings and the employment cost index.

We analyze the contribution of each of the three components of labor market slack to inflation during the expansionary period following the Great Recession. We find that the drop in the on-the-job search rate emerges as the key explanation for why inflation was so low in the U.S. after nine years of economic recovery. Labor misallocation also contributes significantly to keeping inflation persistently subdued following the Great Recession, offsetting the effects of the low unemployment rate.

The surge in labor misallocation right after the recession was due to the exceptionally high stock of unemployed workers who took a first step back onto the job ladder. As a result of the persistent decline in the on-the-job search rate throughout the recovery, the speed at which workers moved to better jobs fell, exacerbating labor misallocation and exerting persistent downward pressures on wages and inflation. Indeed, our model predicts that after nine years of expansion, a significant fraction of the employed workers is still stuck in suboptimal jobs. This prediction is consistent with the micro evidence from the SCE, which shows that in 2017, after eight years of economic recovery, about 30% of workers were not fully satisfied with how their current jobs fit their experience and skills. This persistent rise in bad jobs also accords well with the findings in Autor (2010), Brynjolfsson and McAfee (2011), and Jaimovich and Siu (2018), who show that job polarization intensified following the Great Recession. Jaimovich et al. (2020) document that the deterioration of employment prospects for a large share of workers who were employed in routinary occupations, led to widespread discouragement in searching for jobs.

When we extend the evaluation of our theory to include the periods before the post-Great Recession recovery, we find that our measure of labor market slack performs better in explaining inflation than other popular measures in the literature—such as the labor share of income (as

in Galí and Gertler 1999); the unemployment gap based on the nonaccelerating inflation rate of unemployment (NAIRU); and the hours worked, which features prominently in estimated dynamic general equilibrium models as the key observable variable informing the output gap (e.g., Christiano, Eichenbaum, and Evans 2005). The appalling performance of these traditional measures of slack in explaining inflation over the last decade largely drives this result. In contrast, our measure of slack does a good job at explaining inflation across the three business cycles of this longer sample period (1990-2018).

The assumption that the on-the-job search rate varies stochastically over time is meant to capture all those cyclical factors that drive the decision to search on the job, as well as compositional changes in the propensity to search within the pool of employed workers.[2] At first, we do not explicitly model these compositional changes and assume that on-the job search is exogenous. One advantage of this approach, is that it allows us to derive a closed-form expression for labor market slack, which can be measured from the rate of unemployment and the EE transition rate, without solving the model. A disadvantage, is that does not capture the incentives that workers may have to search more actively when employed in a bad match or during expansions, when getting job offers is easier. These effects may reduce the fall of the on-the-job search rate in recent years, which is key to explain the missing inflation. We show that endogenizing the rate of on-the-job search does not materially affect our conclusions on the dynamics of labor market slack and on the missing inflation in the previous decade. This is because the rate of on-the-job search is effectively implied by the joint behavior of unemployment and EE rates, even in this richer version of the model. Interestingly, the stagnant EE flows during the past economic recovery is largely explained by the decline in the search rate of those workers who could not secure a good offer and thereby remained stuck in a bad match, corroborating the discouragement hypothesis.

Our model features an occasionally binding zero lower bound (ZLB) constraint on the nominal interest rates. Introducing this constraint is important given that the severity of the Great Recession, which in our analysis is captured by the sharp increase of the unemployment rate in 2008 and 2009, drives the current and expected nominal interest rates to the ZLB for several months in our model. We develop an innovative method to solve and simulate models when the ZLB constraint is binding. Our method does not rely on assuming perfect foresight.

Moscarini and Postel-Vinay (2019) pioneer a New Keynesian model in which cyclical labor misallocation brings about deflationary pressures. In building our model, we draw from their groundbreaking contribution. These scholars use the model to show that the degree of labor misallocation is a better predictor of inflation than the rate of unemployment. Our contribution differs from that of Moscarini and Postel-Vinay (2019) in two important respects. First, while

[2]For instance, research shows that workers who get hired at the beginning of an expansion tend to be more eager to search on the job and to climb up the ladder than those less dynamic workers who generally find jobs only after many years of economic expansion (Cahuc, Postel-Vinay, and Robin 2006).

their empirical analysis is reduced form and external to their structural model, we take our structural model to the data using time series methods. Second, while Moscarini and Postel-Vinay focus exclusively on the role of cyclical labor misallocation, we emphasize the importance of the propensity to search on the job for the dynamics of wages and inflation. We show that this propensity can be measured using aggregate labor market flows and the macro estimates are validated using micro data. Crucially, allowing the on-the-job search rate to vary over time is key to explaining the missing inflation of the past decade. When the search rate is constant, the acceptance ratio, which is the ratio of EE to UE flow rates, is a leading indicator for inflationary pressures. This ratio is a proxy for the degree of cyclical labor misallocation and a low value of this ratio predicts high inflation.[3] Through the end of 2019, the acceptance ratio was lower than its pre-Great Recession average in the data (see Appendix A). Our model jointly explains this low acceptance ratio, the persistent increase in bad jobs, and the low inflation in the most recent years with the decline in the incidence of on-the-job search. According to our model the acceptance ratio is currently low in the data, not because employment is efficiently allocated but because fewer workers are searching on the job.

Understanding the search behavior of the employed using disaggregated labor data is an active area of ongoing research, to which we will refer in the paper. In this paper, we stress the importance of this line of research to improve our understanding of inflation. Abraham and Haltiwanger (2019) survey this literature and analyze the behavior of a measure of labor market tightness extended to include all effective job seekers, including employed workers searching on the job. A fall in the rate of on-the-job search, in their view, reduces the number of job seekers and thereby increases labor market tightness, which in turn puts upward pressure on wage and price inflation. In our model, a fall in the rate of on-the-job search instead reduces interfirm wage competition, thereby inducing lower inflationary pressures.

A voluminous literature has shown that inflation has become less sensitive to changes in the traditional measures of labor market slack since the early 1990s, flattening out the slope of the estimated price Phillips curve (Atkenson and Ohanian 2001; Stock and Watson 2007, 2008, and 2019).[4] Del Negro et al. (2020) show that the comovement of unemployment and the labor share over the business cycle is stable over time. Since the labor share is a proxy for marginal costs in standard New Keynesian models, these scholars interpret the stability of this comovement as evidence of a flattening in the Phillips curve. We note that in our model, the presence of on-the-job search breaks the traditional link between marginal costs and the labor share, by connecting marginal costs to the current and expected EE rate. So our finding that

[3]The fraction of accepted offers is lower when more workers are employed in high-productivity jobs. If workers are efficiently allocated, outside offers are declined and matched by the current employer, raising production costs and inflation.

[4]McLeay and Tenreyro (2019) provide a intriguing theoretical reason for why the Phillips curve has become flatter in recent years.

reduced on-the-job search accounts for a fall in marginal costs and inflation at a time of low unemployment is not at odds with the finding of a stable relationship between unemployment and the labor share.

The paper is organized as follows. In Section 2, we provide the motivation for our paper by laying out the missing inflation puzzle. The model from which we derive the novel measure of labor market slack is introduced in Section 3. We explain the empirical strategy and results in Section 4. We check the robustness of our results in Section 5, where we study a model in which workers optimally chose whether to search or not to search on the job. In Section 6 we discuss the performance of the proposed measure of slack in fitting inflation on a longer sample starting in the early 1990s. In Section 7, we present our conclusions.

2 The Missing Inflation Puzzle

The New Keynesian model is the most popular framework to study inflation. A key building block of this framework is the New Keynesian Phillips curve, which posits that inflation π_t hinges on the expected dynamics of future real marginal costs φ_t:

$$\pi_t = \kappa\varphi_t + \beta E\pi_{t+1}, \tag{1}$$

where κ denotes the slope of the curve and β the discount factor. In empirical applications, the real marginal cost φ_t is proxied in a variety of ways. We consider proxies related to the following three traditional theories of the Phillips curve: (*i*) old-fashioned theories (recently revived by Galí, Smets, and Wouters 2011) that link inflation to the current and expected unemployment gap; (*ii*) the standard New Keynesian theory (derived from models with no labor frictions), which suggests that the labor share alone is the key determinant of the inflation rate (Galí and Gertler 2000); (*iii*) a variant of the standard New Keynesian theory, based on models that account for search and matching frictions, which explains inflation using current and expected measures of the labor share as well as UE flow rates (Krause, Lopez-Salido, and Lubik, 2008).[5]

While there are more sophisticated versions of the New Keynesian Phillips curve, which, for instance, feature price indexation, we focus here on the simpler version of this curve to facilitate comparability with the model presented in the next section. We discuss the extension to the case of price indexation in Appendix C and show that it does not affect our main conclusions.

By solving equation (1) forward, we can express expected inflation as the sum of the current and future expected real marginal costs. We estimate a Vector Autoregression (VAR) model to forecast the future stream of the three aforementioned measures of real marginal costs. The forecasts of real marginal costs are launched from every quarter during the post-Great Recession

[5] To make the paper self-contained, we summarize how this third series of marginal costs is constructed in Appendix B. We refer the interested reader to Krause, Lopez-Salido, and Lubik (2008) for more details.

recovery and are then plugged into the Phillips curve, which returns the predicted inflation rate by each of the three theories of marginal costs in every quarter of the recovery. To conduct this exercise, we set the discount factor β to 0.99 (data are quarterly) and a slope of the Phillips curve κ equal to 0.005, so as to fit inflation at the beginning of the post-Great Recession recovery (2009–2011). The resulting Phillips curve is fairly flat and in line with estimates obtained by Del Negro et al. (2020) for the U.S economy over the post-1990 period, using standard measures of slack. While the slope of the Phillips curve affects the magnitude of inflation predicted by the three measures of slacks, it does not affect the point in time when inflation rises above its long-run level, which is what we are interested in (see Appendix C). A flatter Phillips curve in and of itself does not explain why inflation has remained subdued for an entire decade.

To estimate the VAR model, we use the following observable variables: the labor share, the job finding rate, real wages, the civilian unemployment rate, real gross domestic product (GDP), real consumption, real investment, inflation according to the Consumer Price Index (CPI), and the federal funds rate (FFR).[6] We detrend the observable variables by using their 8-year past moving average trend. The only exception is when we construct the unemployment gap, for which we use the short-term NAIRU estimates.[7] We rely on the NAIRU estimates to construct the unemployment gap as this practice is very popular in those studies whose object is to estimate the Phillips curve. The sample period for estimation is from 1958Q4 through 2017Q4.

Figure 2 shows that all the three traditional theories of marginal costs predict that inflation should have been above its long-run level (positive inflation gap) by the end of 2012. None of these theories is able to account for why inflation stayed so low for so many years after the Great Recession because all three proxies for marginal costs improved quickly in the first years of the economic recovery. Consequently, the VAR model's forecasts of future marginal costs go up at a relatively early stage of the recovery, which leads the three New Keynesian Phillips curves to predict inflation above its long-run level. As shown in Appendix E, a state-of-the-art structural model, such as the model studied in Smets and Wouters (2007), also fails to explain the missing inflation.

It is worth noting that this VAR approach is general and agnostic because we do not impose a requirement that expectations about future labor costs are formed according to the Phillips curve. Imposing such a restriction on the VAR structure may lead to misspecification that would most likely worsen the quality of the forecasts of real marginal costs. That said, our main conclusions are not affected by imposing this restriction, and our approach is more appealing in that the unrestricted VAR model is a reduced-form, theory-free representation for the data that is less prone to misspecification than structural theory-based models. Another advantage

[6] Details on how these series are constructed are in Appendix D.

[7] Using the long-term NAIRU would not change our main conclusions.

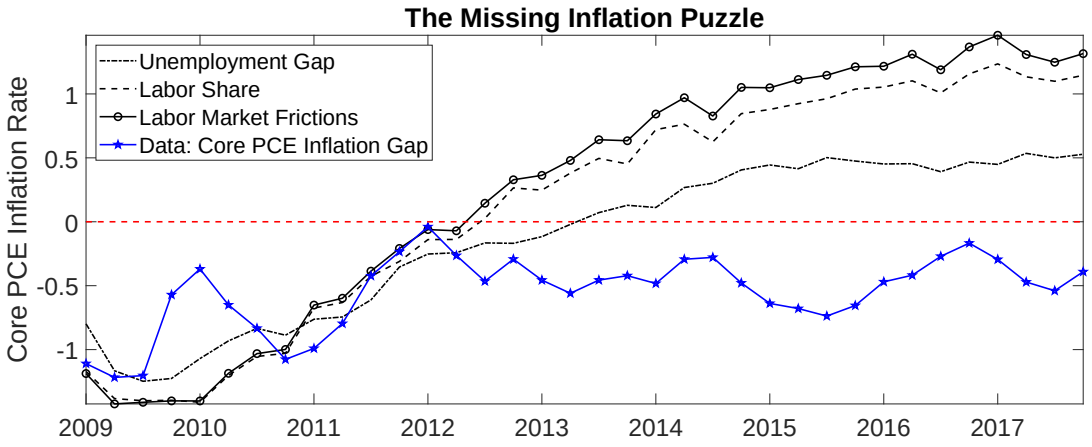

Figure 2: Core PCE inflation gap from 2009Q1 through 2017Q3 and inflation dynamics predicted using three traditional theories of inflation.

of our approach is that VAR models generally provide reliable macroeconomic forecasts.[8]

3 A General Equilibrium Model with On-the-Job Search

The failure of the traditional measures of labor market slack to explain inflation in the past decade motivates the need of an alternative concept of slack. We build this concept using a New Keynesian model in which a time-varying fraction of workers search on the job and firms have to compete to attract or retain these workers by bidding up their wage offers.[9]

3.1 The Economy

The economy is populated by a representative, infinitely lived household, whose members' labor market status is either unemployed or employed. All members of the household are assumed to pool their income at the end of each period and thereby consume the same amount. The labor market is frictional and workers search for jobs whether they are unemployed or employed. While all unemployed workers are also job seekers, it is assumed that any employed worker can search in a given period with a probability s_t, which is assumed to follow an exogenous first-order autoregressive AR(1) process with Gaussian shocks. Time variation in s_t is meant

[8]For Bayesian VAR models to deliver reliable macro forecasts, the choice of the prior is key. We check that VAR forecasts are accurate in sample and follow the conventions established by the forecasting literature. Specifically, we use the unit-root prior introduced by Sims and Zha (1998) and choose the prior hyperparameters, which determine the direction of the Bayesian shrinkage, so as to maximize the marginal likelihood.

[9]Key empirical studies that explicitly allow for search and matching frictions in New Keynesian models include Gertler, Sala, and Trigari (2008), Krause, Lopez-Salido, and Lubik (2008), Ravenna and Walsh (2008) and Christiano, Eichenbaum, and Trabandt (2016). We deviate from these studies by considering the role of on-the-job search and by focusing on inflation. Gertler, Huckfeldt, and Trigari (2019) develop a model where productivity is match-specific, and workers climb the ladder by searching on the job. Their paper abstracts from nominal rigidities and focuses on the wage cyclicality of the newly hired workers.

to capture all those cyclical factors that are responsible for changes in the average rate of on-the-job search in the data, including compositional changes in the propensity to search in the pool of employed workers. Households trade one-period-government bonds B_t.

We distinguish two types of firms: labor-service producers and price setters. The service sector comprises an endogenous measure of worker–firm pairs that match in a frictional labor market and produce a homogeneous nonstorable good. Productivity $y \in \{y_g, y_b\}$ is match-specific and can be either good or bad, with $y_g > y_b > 0$. We let ξ_g denote the probability that upon matching the productivity draw is good and $\xi_b = 1 - \xi_g$ the probability that the draw is bad. The output of the match is sold to price-setting firms in a competitive market at the relative price φ_t (the price of the labor service relative to that of the numeraire), and transformed into a differentiated product. Specifically, one unit of the service is transformed by firm i into one unit of a differentiated good $y_t(i)$. These firms set the price of their goods subject to Calvo price rigidities. Households consume a bundle C_t of such varieties in order to minimize expenditure. This bundle is the numeraire for this economy and its price is denoted by P_t. The monetary authority sets the nominal interest rate of the one-period government bond following a Taylor rule subject to a nonnegativity constraint. The fiscal authority levies lump-sum taxes T_t to finance maturing government bonds.

3.2 The Labor Market

The labor market is frictional and governed by a meeting function that brings together vacancies and job seekers. The pool of workers looking for jobs at each period of time t is given by the measure of workers who are unemployed at the beginning of a period, $u_{0,t}$ plus a fraction s_t of the workers who are employed, $1 - u_{0,t}$. Denoting the aggregate mass of vacancies by v_t, we can define labor market tightness as

$$\theta_t = \frac{v_t}{u_{0,t} + s_t(1 - u_{0,t})}. \tag{2}$$

We assume that the meeting function is homothetic, which implies that the rate at which searching workers find a vacancy, $\phi(\theta) \in [0,1]$, and the rate at which vacancies draw job seekers, $\phi(\theta)/\theta \in [0,1]$, depend exclusively on θ and are such that $d\phi(\theta)/d\theta > 0$ and $d[\phi(\theta)/\theta]/d\theta < 0$.

Because of frictions in the labor market, wages deviate from the competitive solution. It is assumed that wage bargaining follows the sequential auction protocol of Postel-Vinay and Robin (2002). Namely, the outcome of the bargaining is a wage contract, i.e., a sequence of state-contingent wages, which promises to pay a given utility payoff in expected present value terms (accounting also for expected utility from future spells of unemployment and wages paid by future employers). The commitment of the worker–firm pair to the contract is limited, in the sense that either party can unilaterally break up the match if either the present value of firm

profits becomes negative or the present value utility from being employed falls below the value of being unemployed. The contract can be renegotiated only by mutual consent: if an employed worker meets a vacancy, the current and the prospective employer observe first the productivity associated with both matches, and then engage in Bertrand competition over contracts. The worker chooses the contract that delivers the largest value.

The within-period timing of actions is as follows: all the unemployed workers and a fraction s_t of the employed search for a job at the beginning of the period. Next, some workers move out of the unemployment pool, while successful on-the-job seekers have their wage renegotiated and possibly move up the ladder. Then production takes place and wages are paid. This timing implies that workers who are unemployed at the beginning of the period can produce at the end of the same period if they find a job. And similarly, workers who are employed at the beginning of the period may be producing in a different job at the end of the same period if they switch employers. Finally, a fraction δ of the existing matches is destroyed.

These assumptions imply the following dynamics for the aggregate state of unemployment. Denote the stock of end-of-period employed workers as

$$n_t = 1 - u_t. \tag{3}$$

Aggregate unemployment at the beginning of a period is given by

$$u_{0,t} = u_{t-1} + \delta n_{t-1}, \tag{4}$$

while aggregate unemployment at the end of a period is given by

$$u_t = u_{0,t} \left[1 - \phi\left(\theta_t\right)\right]. \tag{5}$$

3.3 Households

Households solve two problems. First, they decide how to optimally allocate their consumption of the aggregate good over time. Second, they solve an intratemporal problem to optimally choose the composition of the aggregate good in terms of differentiated goods sold by the price setters. All workers share their consumption risk within the households, allowing us to solve the problems from the perspective of a representative household.

The intertemporal maximization problem The representative household enjoys utility from the consumption basket C_t and from the fraction of its members who are not working and are therefore free to enjoy leisure. The parameter b controls the marginal utility of leisure. We assume that the utility function is logarithmic in consumption and let μ_t denote the preference shock to consumption, which is assumed to follow a Gaussian AR(1) stochastic process in logs.

The resources available to consume at a given point in time t include government bond holdings B_t; profits from the price setters, which sell differentiated goods, D_t^P; profits from the service firms D_t^S; wages from the workers who are employed; and transfers from the government T_t.

We assume that all unemployed workers look for jobs, and restrict our attention to equilibria where the value of being employed for any worker is no less than the value of being unemployed. In this setup, the measure of workers who are employed is not a choice variable of the household, but is driven by aggregate labor market conditions through the job finding probability $\phi(\theta_t)$. Let $e_t(j) \in \{0,1\}$ be an indicator function which takes the value of one if a worker j is employed after worker reallocation takes place, but before the current-period exogenous separation occurs with probability, δ, and zero otherwise. The intertemporal maximization problem is

$$\max_{\{C_t, B_{t+1}\}} E_0 \sum_{t=0}^{\infty} \beta^t \left[\mu_t \ln C_t + b \int_0^1 (1 - e_t(j))\, dj \right],$$

subject to the budget constraint,

$$P_t C_t + \frac{B_{t+1}}{1+R_t} \leq B_t + \int_0^1 e_t(j)\, w_t(j) + D_t^P + D_t^S + T_t,$$

and the stochastic process for the employment status,

$$\begin{aligned} prob\{e_{t+1}(j) = 1 \mid e_t(j)\} &= e_t(j)\left[(1-\delta) + \delta\phi(\theta_{t+1})\right] + \left[1 - e_t(j)\right]\phi(\theta_{t+1}) \\ prob\{e_{t+1}(j) = 0 \mid e_t(j)\} &= 1 - prob\{e_{t+1}(j) = 1 \mid e_t(j)\}, \end{aligned} \tag{6}$$

and for equilibrium wages $w_t(j)$.[10]

Equation (6) implies that a worker j who is registered as unemployed at the production stage of period t, i.e., $e_t(j) = 0$, will only have a chance to look for jobs at the beginning of next period, and get one with probability $\phi(\theta_{t+1})$. Moreover, a worker employed at time t, i.e., $e_t(j) = 1$, will also be in employment at $t+1$ if she does not separate from the current job between periods at the exogenous rate δ, or if she separates but manages to find a new job with probability $\phi(\theta_{t+1})$ in the next period.

[10] The evolution of individual wages must obey the wage contract negotiated by the worker–firm pair. In these negotiations, workers and firms agree on a present discounted value of the future stream of utility, as we will show later. However, there are many streams of wages that can deliver the promised present discounted value of utility, making the distribution of the individual wages indeterminate. It can be shown that this indeterminacy is inconsequential for aggregate equilibrium outcomes. Nevertheless, as we will clarify later, the real marginal cost, which is the price of the labor service and hence a measure of the average cost of labor, is determined, even though the underyling wage distribution is not.

The intratemporal minimization problem conditions Households minimize total expenditure on all differentiated goods,

$$\min_{q_t(i), i\in[0,1]} \int_0^1 p_t(i)\, q_t(i)\, di, \tag{7}$$

subject to the general Kimball (1995) aggregator assumed in Smets and Wouters (2007):

$$\int_0^1 G\left(q_t(i)/Q_t\right) di = 1. \tag{8}$$

The reason why we choose this particular aggregator will be explained in Section 4.1, where we discuss how we calibrate the key parameter of this aggregation technology. As in Dotsey and King (2005), Levin, Lopez-Salido, and Yun (2007), and Lindé and Trabandt (2018), we assume the following strictly concave and increasing function for $G\left(q_t(i)/Q_t\right)$:

$$G\left(q_t(i)/Q_t\right) = \frac{\omega^k}{1+\varkappa}\left[(1+\varkappa)\frac{q_t(i)}{Q_t} - \varkappa\right]^{\frac{1}{\omega^k}} + 1 - \frac{\omega^k}{1+\varkappa}, \tag{9}$$

where $\omega^k = \frac{\chi(1+\varkappa)}{1+\varkappa\chi}$, $\varkappa \leq 0$ is a parameter that governs the degree of curvature of the demand curve for the differentiated goods and χ captures the gross markup.

The solution of this expenditure minimization problem is the demand function for the differentiated good (i):

$$\frac{q_t(i)}{Q_t} = \frac{1}{1+\varkappa}\left(\frac{P_t(i)}{P_t \Xi_t}\right)^{\iota} + \frac{\varkappa}{1+\varkappa}, \tag{10}$$

where $\varkappa \leq 0$ is a parameter, $\iota = \frac{\chi(1+\varkappa)}{1-\chi}$, Ξ_t is the Lagrange multiplier associated with the constraint (8), and the aggregate price index (i.e., the price of the numeraire) satisfies $1 = \int_0^1 \left(\frac{p_{t,i}}{P_t \Xi_t}\right)^{\frac{\iota}{\omega^k}} di$.

3.4 Price Setters

Price setters buy the (homogeneous) output produced by the service firms in a competitive market at the relative price φ_t, turn it into a differentiated good, and sell it to the households in a monopolistic competitive market. They can re-optimize their price $P_t(i)$ with a period probability $1-\zeta$. If they cannot reoptimize, they adjust their price at the steady-state inflation rate Π. Therefore, the problem of the price setting firm is expressed as follows:

$$\max_{P_{t+s}(i)} E_t \sum_{s=0}^{\infty} \beta^{t+s} \zeta^s \frac{\lambda_{t+s}}{\lambda_t}\left(P_t(i)\Pi^s - P_{t+s}\varphi_{t+s}\right) q_{t+s}(i), \tag{11}$$

subject to the demand function (10). Log-linearization and standard manipulations of the resulting price-setting equation lead to the purely forward-looking New Keynesian Phillips curve, which was shown in equation (1).

As standard in New Keynesian models, the Calvo lottery makes this price-setting problem dynamic; i.e., price setters that are allowed to re-optimize their price at time t find it optimal to forecast the future stream of real marginal costs $\{\varphi_\tau\}_{\tau=t}^{\infty}$. This is because price setters anticipate that they may not be able to re-optimize their price in the next periods. In our model, the price setters' real marginal costs φ_t coincide with the relative price of the labor service, and hence, the optimizing price setters care about the determinants of that price, which are the focus of the next section.

3.5 Service Sector Firms: Free-Entry Condition

In this section, we introduce the free-entry condition to the labor service firm and discuss the pivotal role played by this condition in determining the dynamics of price setters' marginal costs and inflation in the model. This condition implies that entrant firms will make zero profits in expectations; i.e., expected costs are equal to the expected surplus after the match is formed. We first discuss the expected costs incurred by entrant service sector firms and then the expected surplus.

Service firms have to pay an advertising cost c per period. In addition, to form a match and produce, they also have to pay a sunk fixed cost of hiring c^f. The expected cost of creating a job equals $c^f + \frac{c}{\varpi_t}$, where ϖ_t is the vacancy filling rate and ϖ^{-1} measures the expected number of periods that is required to meet a worker.

The expected return from a match depends on whether the worker matched is employed or unemployed. Following Postel-Vinay and Robin (2002) and Moscarini and Postel-Vinay (2018, 2019), it is assumed that unemployed workers have no bargaining power, so the firm will appropriate the entire surplus of the match, which will in turn depend on its quality. If the firm meets an employed worker instead, the firm engages in Bertrand competition with the incumbent firm in an attempt to poach the worker away from the current match. An important implication of assuming Bertrand competition is that an increase in wages is not necessarily backed by a rise in workers' productivity. This can happen, for instance, when a worker renegotiates upward the value of a contract, as their employer agrees to match the offer of a poaching firm. This temporary decoupling between wages and the worker's productivity is key for the job ladder to have meaningful implications for inflation. As we will show, these assumptions also imply that the worker's ability of extracting more and more surplus from a match depends on her position on the job ladder.

While the assumption that unemployed workers have no bargaining power is undoubtedly stark, it provides tractability, allowing for an analytical characterization of the expected sur-

pluses that appear in the free-entry condition. Such an analytical characterization turns out to be very useful in providing intuition about the link between the labor market and inflation in the model, which will be the focus of Section 3.6.

Importantly, the Postel-Vinay and Robin's bargaining protocol breaks the link between labor market tightness and wages, allowing us to isolate the effects of searching on the job on firms' wage competition. Specifically, a drop in the on-the-job search rate leads to a fall in the share of job seekers that are employed, thereby reducing interfirm wage competition and the cost of labor service. This is the first paper that focuses on this channel to explain inflation. Indeed, in a standard New Keynesian model with search and matching, a fall in the rate of on-the-job search reduces the number of job seekers, increasing labor market tightness and wages. For instance, this channel is emphasized in Abraham and Haltiwanger (2019) and in most of the reduced-form labor literature reviewed in that paper. Our model-based empirical analysis is therefore constructed to derive a simple measure of slack that isolates the role of interfirm wage competition and investigate how far it can go in explaining the recent dynamics of inflation.

To illustrate how Bertrand competition works in our model, let y and y' denote the match quality with the incumbent and the poaching firm, respectively. We distinguish three possible contingencies.

1. $y = y_g$ and $y' = y_b$. In this case the poaching firm is a worse match for the worker. Bertrand competition implies that the incumbent firm will retain the worker and poaching is not successful. If the worker was hired from the state of unemployment, she appropriates the surplus $S_t(y_b)$ because Bertrand competition forces the incumbent to pay the worker the highest value the poaching firm is willing to pay her. If the worker was not hired from a state of unemployment, there is no change in the value of her contract.

2. $y = y'$ for $y \in \{y_b, y_g\}$. Match quality is the same for the two firms, and the worker will be indifferent between switching jobs or staying. We assume that switching takes place with probability ν (a nonzero value for this parameter is required to match the high churning rate in the U.S. labor market when calibrating the model's steady-state parameters). In either case, the firm that ends up with the worker relinquishes all the surplus $S_t(y)$.

3. $y = y_b$ and $y' = y_g$. Match quality is lower with the incumbent firm, so the worker is poached. Bertrand competition implies that the worker is given the highest surplus the incumbent firm is willing to pay her, i.e., $S_t(y_b)$. The poaching firm's surplus is therefore the residual value of the match: $S_t(y_g) - S_t(y_b)$.

To sum up, entrant labor service firms can get a nonzero surplus from meeting an employed worker only if the worker is in a bad match and the firm is a good match for the worker. As a

result, the free-entry condition can be written as follows:

$$c^f + \frac{c}{\varpi_t} = \frac{u_{0,t}}{u_{0,t} + s_t(1-u_{0,t})}\left\{\xi_b S_t(y_b) + \xi_g S_t(y_g)\right\} + \frac{s_t(1-u_{0,t})}{u_{0,t} + s_t(1-u_{0,t})}\left\{\xi_g \frac{l^0_{b,t}}{1-u_{0,t}}\left[S_t(y_g) - S_t(y_b)\right]\right\}, \tag{12}$$

where $l^0_{b,t}$ denotes the measure of workers who, at the beginning of period t, are employed in low-quality matches ($l^0_{b,t} + l^0_{g,t} + u_{0,t} = 1$) and s_t is the on-the-job search rate. The term $s_t(1-u_{0,t})$ denotes the measure of employed workers searching on the job at the beginning of period t, and $u_{0,t} + s_t(1-u_{0,t})$ is the measure of all job seekers at the beginning of period t.

The left-hand side is the expected costs of posting a vacancy, which has been discussed above. The expected return from forming a match, on the right-hand side, depends on the employment status, on the quality of the meeting, and, in the case the firm meets an employed worker, also on the quality of the existing match. Three contingencies will give a nonzero surplus to the firm and will hence appear in the right-hand side of the free-entry equation (12). The expected return on the right-hand side is the average of the surplus accrued in these three contingencies weighted by their respective probabilities.

The first contingency is when the entrant firm meets an unemployed job seeker, with probability $u_{0,t}/[u_{0,t} + s_t(1-u_{0,t})]$, and the job seeker is a bad match for the firm, with probability ξ_b. In this case the meeting gives the firm the surplus $S_t(y_b)$. The second contingency is when the entrant firm meets an unemployed job seeker who turns out to be a good match, with probability ξ_g, providing the firm with the surplus $S_t(y_g)$. These two expected returns appear in the first term on the right-hand side of the free-entry equation (12). The third contingency, i.e., the second term in the right-hand side of the free-entry equation (12), occurs when the firm meets an employed worker, with probability $s_t(1-u_{0,t})/[u_{0,t} + s_t(1-u_{0,t})]$, and the following two conditions are met: (i) the worker is a good match for the entrant firm, which occurs with probability ξ_g, and (ii) the worker is currently in a bad match, which happens with probability $\xi_g l^0_{b,t}/(1-u_{0,t})$.[11] As explained above, this is the only case in which an entrant firm can extract a nonzero surplus from meeting with an employed worker.

Moscarini and Postel-Vinay (2019) show that the surplus function can be written as follows

$$S_t(y) = y\mathcal{W}_t - \frac{b\lambda_t^{-1}}{1-\beta(1-\delta)}, \tag{13}$$

[11] Note that $l^0_{b,t}$ denotes the share of workers that are employed in a bad match at the beginning of the period. We rescale this share by the fraction of employed workers at the beginning of the period $(1-u_{0,t})$ so as to obtain the conditional probability of meeting a bad match.

where λ_t is the Lagrange multiplier with respect to the household's budget constraint and

$$\mathcal{W}_t = \varphi_t + \beta (1-\delta) E_t \frac{\lambda_{t+1}}{\lambda_t} \mathcal{W}_{t+1}. \tag{14}$$

See Appendix F for details on the derivations in the context of our model. From the point of view of a labor service firm, $\mathcal{W}_t$ can be interpreted as the expected present discounted value of the entire stream of current and future prices at which the homogeneous good is sold by the firm until separation from the worker.

3.6 Wage Competition, Labor Market Slack, and Inflation

In this section, we conduct a partial-equilibrium thought experiment to build intuition about what are the main drivers of inflation in the model. We will verify this intuition later in Section 4.4. In this thought experiment, we consider the free-entry condition (12) in isolation and assume that either the unemployment rate u_t^0 increases, the fraction of workers in a bad match $l_{b,t}^0$ rises, or employed workers search less frequently (s_t decreases). The direct effect of these changes on the free-entry condition is to increase the expected profits (i.e., the right-hand side of the free-entry condition) because it is more likely for firms to meet a worker they can extract a positive surplus from. As a result, more firms want to enter the labor service sector enticed by the expected gains from posting a vacancy.

Two variables adjust to restore the free-entry condition. On the one hand, the increase in vacancies leads to a decrease in the vacancy filling rate, ϖ_t, and to an increase in the expected cost of entry (i.e., the left-hand side of equation (12)). On the other hand, the expected discounted stream of the relative price of labor service $\mathcal{W}_t$ falls, lowering surpluses $S_t(y)$, as implied by equations (13) and (14), and further dissuading firms from posting new vacancies until the free-entry condition is satisfied. This drop in the relative price of labor service causes price setters' marginal costs to fall, lowering inflation.[12] As it will become clear later, the second channel turns out to be quantitatively important in the calibrated model.

Based on this reasoning, we conjecture that a key variable that affects inflation in the model is the probability that, conditional on a contact, firms entering the labor service sector are not engaged in a wage competition that leads them to relinquish the entire surplus to the worker. This probability is defined as follows:

$$\Sigma_t \equiv \frac{u_{0,t}}{u_{0,t} + s_t (1-u_{0,t})} + \frac{s_t (1-u_{0,t})}{u_{0,t} + s_t (1-u_{0,t})} \xi_g \frac{l_{b,t}^0}{1-u_{0,t}}, \tag{15}$$

[12] The present discounted stream of real marginal costs $\mathcal{W}_t$ falls only in the presence of nominal rigidities. With flexible prices, real marginal costs are constant and the equilibrium of the free-entry condition is restored only through a change in the vacancy filling rate.

where the first term on the right-hand side is the probability of meeting an unemployed worker and the second term is the probability of meeting a worker employed in a bad match, who is searching on the job, and turns out to be a good match for the poaching firm. Equation (15) clarifies that labor market slack is a multidimensional object combining the rate of unemployment, the degree of cyclical misallocation, and the on-the-job search rate.

As explained in the partial-equilibrium thought experiment, a high value of this probability implies a low intensity of wage competition, leading to downward pressures on price setters' marginal costs and inflation. Hence, this probability can be thought of as a measure of labor market slack in this model. While this intuition is obtained using this partial-equilibrium thought experiment, the link between this measure of slack Σ_t and inflation also holds in general equilibrium, as we will show in Section 4.4.

The notion of labor market slack provided in Equation (15) has two main advantages. First, it will allow us to decompose inflation into three main drivers: the unemployment rate $u_{0,t}$, the degree of labor misallocation $l^0_{b,t}$, and the on-the-job search rate s_t. Such a decomposition will turn out to be very useful to illustrate what are the key variables allowing the model to explain the low inflation of the past decade. Second, we do not need to solve the model to quantify this measure of labor market slack. In fact, it can be measured directly from the unemployment rate and the EE flow rate in the data, as we will show in Section 4.3.

3.7 The Dynamic Distribution of Match Types

The laws of motion for bad and good matches are

$$l_{b,t} = \left[1 - s_t \phi\left(\theta_t\right) \xi_g\right] l^0_{b,t} + \phi\left(\theta_t\right) \xi_b u_{0,t}, \tag{16}$$

$$l_{g,t} = l^0_{g,t} + s_t \phi\left(\theta_t\right) \xi_g l^0_{b,t} + \phi\left(\theta_t\right) \xi_g u_{0,t}. \tag{17}$$

In the above equations, we let $l_{b,t}$ and $l_{g,t}$ denote the end-of-period measure of bad and good matches, respectively. We let $l^0_{b,t}$ and $l^0_{g,t}$ denote beginning-of-period values. In turn, $l_{b,t}$ is equal to the sum of the bad matches at the beginning of a period that did not move up the ladder by finding a high-quality match within the period, $\left[1 - s_t \phi\left(\theta_t\right) \xi_g\right] l^0_{b,t}$, plus the new hires from the unemployment pool who turned out to draw a low-quality match, $\phi\left(\theta_t\right) \xi_b u_{0,t}$. Indeed, job-to-job flows from bad- to good-quality matches are given by the fraction of badly matched employed workers, $l^0_{b,t}$, who search on the job with exogenous probability s_t, meet a vacancy with probability $\phi\left(\theta_t\right)$, and draw a good match with probability ξ_g.

The end-of-period measure of good matches is instead given by the beginning-of-period measure of good matches $l^0_{g,t}$, plus the job-to-job inflows from bad matches $s_t \phi\left(\theta_t\right) \xi_g l^0_{b,t}$, and the unemployed hired in a good job, $\phi\left(\theta_t\right) \xi_g u_{0,t}$. Using the identity $l^0_{i,t+1}\left(y\right) = \left(1 - \delta\right) l_{i,t}\left(y\right)$ for $i = \{b, g\}$,we can rewrite the dynamic equations (16) and (17) to express the laws of motion

for bad and good jobs at their beginning-of-period values:

$$l_{b,t+1}^0 = (1-\delta)\left\{\left[1-s_t\phi\left(\theta_t\right)\xi_g\right]l_{b,t}^0+\phi\left(\theta_t\right)\xi_b u_{0,t}\right\}, \tag{18}$$

$$l_{g,t+1}^0 = (1-\delta)\left\{l_{g,t}^0+s_t\phi\left(\theta_t\right)\xi_g l_{b,t}^0+\phi\left(\theta_t\right)\xi_g u_{0,t}\right\}. \tag{19}$$

3.8 Policymakers and Market Clearing

The fiscal authority levies lump-sum taxes to repay its maturing bonds in every period. The monetary authority follows a Taylor rule when the nominal interest rate R_t is not constrained by the zero lower bound:

$$\frac{R_t}{R^*} = \max\left\{\frac{1}{R^*}, \left(\frac{R_{t-1}}{R^*}\right)^{\rho_r}\left[\left(\frac{\Pi_t}{\Pi^*}\right)^{\phi_\pi}\left(\frac{Q_t}{Q^*}\right)^{\phi_y}\right]^{1-\rho_r}\right\}, \tag{20}$$

where $\frac{1}{R^*}$ represents the lower bound of the nominal interest rate, $\rho_r \in [0,1)$ captures the degree of interest rate smoothing, and the parameters $\phi_\pi > 1$ and $\phi_y > 0$ capture how strongly the central bank responds to inflation (in deviation from the target Π^*) and output (in deviation from its potential level Q^*).

We do not include monetary shocks in equation (20) because these shocks cannot be separately identified by preference shocks in our empirical analysis. Indeed, the observables, which are the unemployment rate and the EE flow rate, respond very similarly to these two shocks.[13] To disentangle these two shocks, one has to add some other series—e.g., the nominal interest rate. However, adding nominal variables is undesirable as these variables could indirectly give our model information about the inflation rate. Instead, in our empirical analysis we aim to asses the model's ability to explain inflation in the post-Great Recession recovery using solely real labor market variables. We consider this an important feature of our analysis.

Market clearing in the market of price-setting firms implies that the quantity sold summing over all producers i must be equal to the production in the service sector:

$$y_g l_{g,t} + y_b l_{b,t} = \int_0^1 q_t\left(i\right)di.$$

In turn, aggregate output from price setters must equal aggregate demand from the households:

$$\int_0^1 q_t\left(i\right)di = Q_t\int_0^1\left(\frac{1}{1+\varkappa}\left(\frac{P_t\left(i\right)}{P_t\Xi_t}\right)^{\iota}+\frac{\varkappa}{1+\varkappa}\right)di,$$

where we have made use of the demand function in equation (10). Substituting the profits

[13] We note a fair amount of cannibalization between these two shocks when monetary shocks are added to the analysis. As a result, our main results would not change.

of all firms into the household's budget constraint yields the aggregate resource constraint in Moscarini and Postel-Vinay (2019).

4 Empirical Strategy

In section 4.1 we discuss the calibration strategy, and in Section 4.2 we examine the propagation of the shocks to preferences and search intensity. In Section 4.3, we show how to measure some key labor market variables, such as the degree of cyclical labor misallocation $l^0_{b,t}$; the on-the-job search rate; and our measure of labor market slack Σ_t, which we introduced in Section 3.6, using the observed unemployment rate and the EE flow rate. In Section 4.4, we verify the conjecture about the link between our measure of slack Σ_t and inflation. We evaluate the model's ability to explain inflation in Section 4.5. In Section 4.6, we present the micro evidence on the behavior of the on-the-job search rate and use it to validate the rate measured by using aggregate labor market flows.

4.1 Calibration

We calibrate the steady state of the model to the U.S. economy at monthly frequencies. To do so, we assume a Cobb-Douglas matching function $M_t = \phi_0 \left[u_{0,t} + s_t \left(1 - u_{0,t}\right)\right]^{1-\psi} v_t^{\psi}$, where $\psi \in (0,1)$ is an elasticity parameter and $\phi_0 > 0$ is a scale factor.

The calibration of the steady state requires assigning values to the following 11 parameter values: β, ϕ_0, δ, y_b, y_g, ν, b, ξ_g, c, c^f and s. We set the discount factor β to match an annual real interest rate of 1.5%, which is in line with the median of individual economic projections about the real long-term interest rate from various Federal Reserve's Board members, Federal Open Market Committee (FOMC) members, or FOMC participants (known as *Summary of Economic Projections*, or SEP).[14] We normalize θ to unity, which allows us to pin down the scale factor ϕ_0, so as to match a job finding rate of about 33 percent, which is the average of the job finding rate computed following Shimer (2005) over a span of 25 years (January 1993-December 2018).[15] The job separation rate δ is implied by the Beveridge curve, under the assumption of a steady-state rate of unemployment of 5.5%. Namely, solving the Beveridge curve for $\delta = \frac{\phi_0 u_0}{1-u_0+\phi_0 u_0}$ yields a separation rate of 0.02. The productivity of a bad match is normalized to one, and the productivity in a good match is set to be 8% higher. We regard this productivity differential as conservative, in the light of values that have been assigned in the calibration of other comparable models with on-the-job search. Our targeted wage differential is in line with evidence from Faberman et al. (2019) based on the *Survey of Consumer Expectations*,

[14] We take the average of these projections from the FOMC meeting of May 2012—the first meeting after which the projections were released—through the meeting of December 2019.

[15] Under the assumption of unitary tightness ($\theta = 1$), the job finding rate becomes equals to ϕ_0.

Calibration			
Parameters	Description	Value	Target/source
Parameters that affect the steady state			
β	Discount factor	0.9987	Real rate 1.5%. (FOMC SEP)
ϕ_0	Scale parameter matching fn	0.3284	Job finding rate - Shimer (2005)
δ	Job separation rate	0.0200	Unemployment rate ($100u_{0,t}$) 5.5%
y_b	Productivity bad matches	1.0000	Normalization
y_g	Productivity good matches	1.0800	Faberman et al. (2019)
ν	Prob. of job switching if indifferent	0.5000	
b	Utility of leisure	0.8082	Calibrated
c	Flow cost of vacancy	0.0124	Calibrated
c^f	Fixed cost of hiring	0.4958	Calibrated
s	On-the-job search rate	0.2598	Calibrated
ξ_g	Probability draw good match	0.2800	Calibrated
χ	Markup parameter	1.2000	20% markup
$\varkappa$	Scale param. Kimball aggregator	10.0000	Smets and Wouters (2007)
ζ	Calvo price parameter	0.9250	Quarterly probability is 80%
Π	Steady-state gross inflation rate	1.0017	Net inflation rate of 2% p.a.
ρ_r	Taylor rule smoothing parameter	0.8500	Conventional
ϕ_π	Taylor rule response to inflation	1.8000	Conventional
ϕ_y	Taylor rule response to output	0.2500	Conventional
ψ	Elasticity of matching function	0.5000	Moscarini and Postel-Vinay (2018)
ρ_μ	Autocorrel. preference shock	0.8000	Fixed
$100\sigma_\mu$	St. dev. preference shock	0.5883	Volatility of the unempl. rate
ρ_S	Autocorrel. job search rate	0.9157	Maximum likelihood estimation
$100\sigma_S$	St. dev. of job search rate shocks	2.5510	Maximum likelihood estimation
Variable	Description	Value	Target/source
Steady-state calibration targets			
$\frac{c}{\varpi}/c^f$	Ratio of variable to fixed cost	0.0780	Silva and Toledo (2009)
$EE \equiv \frac{s\phi\left[l_b^0\left(\xi_b\nu+\xi_g\right)+l_g^0\xi_g\nu\right]}{l_b^0+l_g^0}$	EE transition rate	2.58	Pre-Great Recession EE rate (%)
θ	Labor market tightness	1.000	Normalization
$\frac{l_g^0}{l_g^0+l_b^0}$	Employment share in good jobs	0.6800	Employment share at top 10% firms
$\frac{\left(v_t c+c^f\phi_t[u_{0,t}+s_t(1-u_{0,t})]\right)/H}{\varphi}$	Hiring costs over wages	0.6000	Hiring costs equal 2 weeks of wages

Table 1: Calibrated values for model parameters. Notes: FOMC SEP stands for the Federal Open Market Committee's Summary of Economic Projections. EE stands for employment-to-employment.

which shows that wage gains associated with job switching are about 8%, after controlling for observable characteristics of workers and jobs. Moreover, we noticed that assigning higher values would violate the incentive compatibility constraint, which requires that the surplus of bad matches should be positive both in steady state and in all periods of the sample used to run the empirical exercise of Section 4.5. Finally, we set the probability that workers will accept an equally valuable outside offer to be $\nu = 0.5$. This value is large enough to allow the model to match the average EE flow rate in the U.S. economy. In Appendix G, we show that perturbing the value of ν does not materially affect our results.

This leaves us with five parameters to calibrate: the parameter governing the utility of leisure b, the probability of drawing a good match ξ_g, the flow cost of advertising a vacancy c, the fixed cost of hiring c^f, and the parameter governing search intensity s. These are calibrated

in order to match the following: *(i)* A value of expected hiring costs, including both the variable and the fixed cost component, equal to two weeks of wages.[16] *(ii)* A fraction of good jobs in steady state equal to 67%, which is the share of employment for the top 10% U.S. firms by employment size in the year 2000. *(iii)* A normalized value of labor market tightness equal to one. *(iv)* A ratio of total variable costs of hiring to fixed costs $\frac{c}{\varpi}/c^f$ equal to 0.078. This value is the ratio of pre-match recruiting, screening, and interviewing costs to post-match training costs in the U.S., following the analysis of Silva and Toledo (2009)—which is based on the 1982 Employer Opportunity Pilot Project (EOPP), a cross-sectional firm-level survey that contains detailed information on both pre-match and post-match labor turnover costs in the United States.[17] *(v)* A monthly job-to-job transition rate of 2.5841%, which is the average EE rate (spliced using the quit rate as explained in Section 4.3) measured in the pre-Great Recession sample (April 1990 through December 2007). We note that the value of the parameter s implied by the calibration, 0.2598, is very close to the value of 0.22, which corresponds to the fraction of U.S. workers who engage in on-the-job search every month, as measured using survey data by Faberman et al. (2019). We have checked that the value of b implied by the calibration is consistent with a positive surplus for low-quality matches both in steady state and in every month considered in the empirical exercise of Section 4.5.

The calibration of the probability of a good match ξ_g (conditional on receiving a job offer) relies on the empirical strategy in Moscarini and Postel-Vinay (2016), who exploit the notorious correlation between firm size and productivity by assuming that employed workers climb the ladder when moving to larger firms. In Appendix G we show that our main results are not affected by reasonable variations in the probability of meeting a good match ξ_g.

We set the smoothing coefficient of the Taylor rule to the value of 0.85, which corresponds to a coefficient of around 0.65 in quarterly space, and the response parameters to inflation and output to the values of 1.8 and 0.25, respectively. The parameter χ is set to equal 1.2, which implies a 20% price markup. The steady-state gross rate of inflation is set to equal 1.0016, which implies a central bank's inflation target Π^* of 2% inflation annually, in line with the Federal Reserve's stated inflation objective. Finally, we set the elasticity of vacancies in the matching function ψ to equal 0.5 to be consistent with estimates by Moscarini and Postel-Vinay (2018), which account for workers searching on the job.

The slope of the Phillips curve is determined by the scale parameter of the Kimball aggregator $\varkappa$ and the Calvo parameter ζ, which govern the size of price stickiness. The former

[16] The average wage is proxied by the price of the labor service φ. If we set a target higher than two weeks of wages, the deflationary pressures predicted by the model in the past decade become slightly stronger, strengthening the ability of our model to explain the missing inflation. We also observe that the relative importance of labor misallocation as a component of labor market slack decreases as we target higher hiring costs.

[17] Silva and Toledo (2009) indicate in Table 1 (p.80), that the average pre-match recruiting cost costs is \$105.1, while the average post-match training cost amounts to \$1,359.4.

parameter is set to 10 as in Smets and Wouters (2007). The latter is set to 0.925, which in quarterly frequency implies a probability of not being able to reoptimize prices equal to 0.8. We set the Calvo parameter so that the implied slope of the Phillips curve allows the model to fit inflation at the beginning of the post-Great Recession recovery (2009–2011), following the approach used to calibrate the slope of the Phillips curve in Section 2, where we evaluate the ability of the traditional measures of slack to explain the missing inflation of the past decade. The Kimball aggregator allows us to obtain the targeted value for the slope of the Phillips curve without requiring us to assume an implausibly large degree of price stickiness. Indeed, in the early stages of the recovery, the combined effect of the fall in the on-the-job search rate, the binding ZLB constraint, and the persistent negative demand shocks that caused the Great Recession is to lower the inflation rate predicted by the model by a fair amount, requiring a relatively flat Phillips curve to fit the level of inflation in the data. This shortcoming is typical of small-scale dynamic equilibrium models. It is worth noting that while the slope of the Phillips curve affects the magnitude of inflation predicted by the model, it can be shown to have negligible effects on the point in time when the model predicts inflation to rise above its long-run level. Therefore, the ability of our model to explain the missing inflation of the last decade is not affected by varying the slope of the Phillips curve.

As we will show in Section 4.3, we can use the observed unemployment rate and the EE flow rate in combination with a subset of model equations to obtain the time series of the on-the-job search rate. This series can be retrieved from the data with no need to solve the model. To pin down this series, we just have to take a stand on a few steady-state parameters (e.g., the steady-state job finding rate ϕ and the separation rate δ), which we calibrate using the values shown in Table 1. We use this series to estimate the persistence parameter ρ_S and the standard deviation σ_S via maximum likelihood.

Turning to the parameters affecting the persistence and the volatility of the preference shock, we set the autocorrelation parameter ρ_μ to 0.80 and then we calibrate the standard deviation σ_μ so that the model can match the volatility of the observed unemployment rate in the data (April 1990 through December 2018).[18] The value of the autocorrelation parameter is a bit lower than what is needed to fit the persistence in the U.S. civilian unemployment rate. However, a persistence higher than 0.8 would make this shock propagate as a supply shock moving the unemployment rate and inflation in the same direction.[19] Because the other shock (i.e., the shock to the on-the-job search rate) propagates as a supply shock, the model would lack a demand shock to explain periods in which inflation and the unemployment rate

[18] We pick the unemployment rate as a target variable because it will be used in our main empirical exercise.

[19] If a negative preference shock is very persistent, the fall in vacancy creation becomes so large that it generates a sharp and prolonged contraction in the supply of the service, which in turn implies a persistent increase in its price, i.e., the real marginal cost φ_t. Moreover, the rise in current and future expected marginal costs entails a rise in the rate of inflation, together with a contraction in aggregate production.

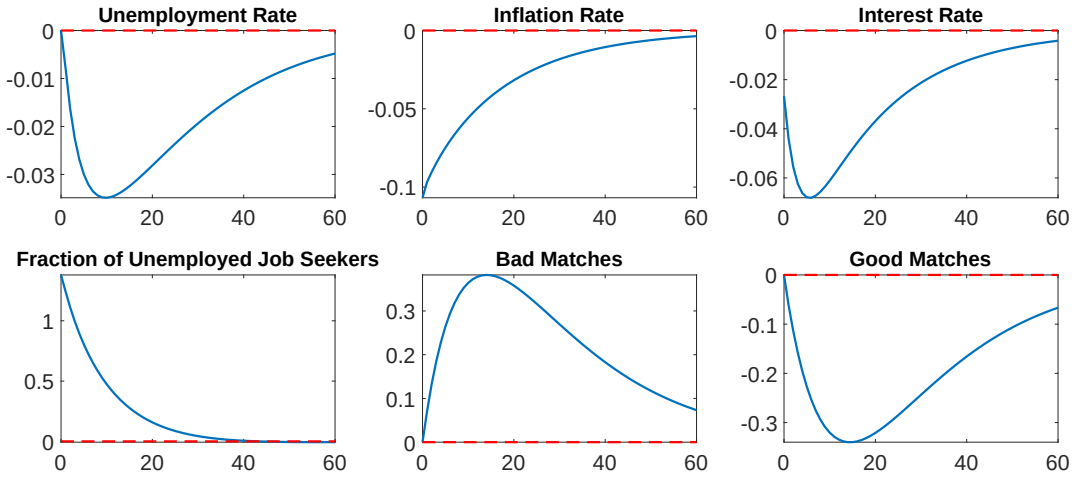

Figure 3: Impulse responses to a shock that lowers the on-the-job search rate by one standard deviation. The unemployment rate, the fraction of unemployed job seekers, and the shares of good and bad matches are measured at the beginning of the period, consistently with the definition of labor market slack Σ_t in equation (15). Units: percentage points. Inflation and interest rate are expressed in annualized rates.

negatively comove.

Model Solution with the Zero Lower Bound (ZLB) Constraint The model is log-linearized around its steady-state equilibrium.[20] However, the zero lower bound introduces a nonlinearity that prevents us from solving the model with standard solution methods. We develop a novel method to find the certainty-equivalence solution to these temporarily non-linear dynamics. Our method does not require us to assume that agents in the model have perfect foresight. Agents update their rational expectations about the duration of the zero lower bound over time after having observed past and current shocks.

Our method relies on appending a sequence of anticipated shocks (dummy shocks) to the unconstrained Taylor rule. Anticipated shocks are known by agents in the current period, but these shocks will hit the economy in future periods. The sequence of these shocks is computed so as to ensure that agents expect that the zero lower bound constraint will be satisfied for the next 36 months in every period.[21] When the constraint is not expected to become binding, these anticipated shocks are set to zero. Obviously, these shocks will have an effect on the expected duration of the ZLB and hence on equilibrium outcomes, requiring us to solve a fixed-point problem, which is described in greater detail in Appendix H. This fixed-point problem does not turn out to be time consuming or computationally challenging in practice.

[20] Rates and shares are linearized; all the other variables are log-linearized.

[21] In none of the periods of our sample, the zero lower bound constraint binds for more than 36 months in expectation. If it did, we would need to add more anticipated shocks to the Taylor rule so as to cover a horizon longer than 36 months.

4.2 Impulse Responses

In this section, we discuss the propagation of the two shocks of the model: the preference shock and the shock to the rate of on-the-job search. We start with the latter shock, whose propagation has been informally discussed in the partial-equilibrium thought experiment of Section 3.6. Figure 3 shows that a fall in the rate at which workers search on the job raises the fraction of job seekers that are unemployed (i.e., the first term on the right-hand side of equation (15) defining labor market slack in our model), lowering the intensity of wage competition and increasing slack. In expectation, producing labor service becomes cheaper for an entrant firm as the likelihood of extracting a positive surplus from meeting a worker increases.

The stock of bad matches rises and the stock of good matches drops. This is quite mechanical, as this shock directly reduces the flow from bad to good jobs, slowing down the allocative mechanism of the ladder. This increase in labor misallocation implies that wage competition is less likely to entirely wipe out the firms' share of surplus and, as a result, the second term of the right-hand side of equation (15) rises, implying a further decline in the intensity of wage competition among firms and a further increase in slack. As a result, inflation drops and the central bank cuts the interest rate, stimulating aggregate demand and reducing unemployment. Moreover, attracted by the expectation of cheaper labor, more firms enter the labor service sector, i.e., more vacancies are created, expanding aggregate supply, which also contributes to lowering the unemployment rate.

Note that the fall in the unemployment rate, in isolation, contributes to lowering the probability for an entrant firm to meet an unemployed worker and hence causes wage competition to become more intense. Yet, as shown in the lower left panel of Figure 3, it turns out that in equilibrium this effect is dominated by the fall in the rate of on-the-job search, which operates in the opposite direction, raising the fraction of unemployed job seekers.

By showing the response of the fraction of job seekers who are unemployed and that of the stock of workers employed in bad matches, we want to provide a decomposition of our measure of slack, Σ_t, defined in equation (15). While in the immediate aftermath of the shock, inflation responds mostly to the rise in the fraction of unemployed job seekers, the persistent change in the match composition of the employment pool weighs down on inflation later on, contributing to keeping inflation below its long-run value for some time. Interestingly, a negative shock to the rate of on-the-job search can generate simultaneously a persistent rise in output, together with a fall in unemployment, inflation, and productivity. Incidentally, these patterns seem to accord well with the dynamics that have characterized the U.S. economy in recent years.

A negative preference shock raises the unemployment rate and cyclical misallocation and lowers inflation. In analogy with the case of a shock to the rate of on-the-job search, in the immediate aftermath of a preference shock the dynamics of inflation reflect mostly the response of the fraction of unemployed job seekers, which influences the intensity of interfirm wage

competition. But as the preference shock dies out, the effect of labor misallocation on labor market slack takes over, inducing inflation to remain persistently below steady state. We show the impulse responses to preference shocks and provide more details in Appendix I.

4.3 The On-the-Job Search Rate in the Macro Data

We show that for a given value of bad and good matches at the beginning of our sample period (i.e., in April 1990), observing the unemployment rate and the EE rate implies the entire time series of the on-the-job search rate s_t, as well as the time series of bad matches $l^0_{b,t+1}$ and good matches $l^0_{g,t+1}$. The exact identification of these variables comes from a set of model's equations and does not require solving the model. We first show this property of the model. Then we use the observed series of the unemployment rate and the EE rate to actually recover the on-the-job search rate and the share of bad matches. We use the equations linearized around the steady-state equilibrium where $\tilde{}$ denotes linearized variables.

The observed series of unemployment rates informs $u_{0,t+1}$ and hence the aggregate unemployment at the end of the period u_t through the following equation

$$\tilde{u}_t = \frac{\tilde{u}_{0,t+1}}{1-\delta}, \tag{21}$$

which is is obtained by combining equations (3) and (4) and linearizing.

Endowed with the end-of-period unemployment rate $\tilde{u}_t$, we can linearize equation (5) to pin down the job finding rate $\tilde{\phi}_t$ at time t as follows:

$$\tilde{\phi}_t = \frac{(1-\phi)\,\tilde{u}_{0,t} - \tilde{u}_t}{u_0}, \tag{22}$$

where u_0 denotes the unemployment rate at the beginning of the period in steady state and ϕ is the job finding rate in steady state. We iterate on equations (21) and (22) using the observed series of the unemployment rate, which yields a time series for the job finding rate $\tilde{\phi}_t$.

We then linearize the definition of the EE flow rate (EE_t) in the model, which reads as follows:

$$EE_t \equiv \frac{s_t \phi(\theta_t)\left[l^0_{b,t}\left(\xi_b \nu + \xi_g\right) + l^0_{g,t}\xi_g \nu\right]}{l^0_{b,t} + l^0_{g,t}}. \tag{23}$$

The EE rate is the ratio of how many workers employed at the beginning of the period have switched jobs (the EE flows) to the total numbers of workers employed at the beginning of the period. Consistently with our model, the EE flows are given by the sum of all the workers who find a better match and the fraction ν of those workers who find an equally valuable match.

Linearizing equation (23) yields the following equation, which expresses the on-the-job search rate $\tilde{s}_t$ as a function of the observed EE flow rate $\widetilde{EE}_t$, the job finding rate $\tilde{\phi}_t$, and a bunch

of variables that are predetermined at time t, such as the distribution of the quality of the matches at the beginning of period t ($\tilde{l}^0_{b,t}$ and $\tilde{l}^0_{g,t}$):

$$\begin{aligned}\tilde{s}_t &= \frac{s}{EE}\widetilde{EE}_t - \frac{s}{\phi}\tilde{\phi}_t - \frac{s}{\nu\left(l^0_b + l^0_g\right)}\left[\frac{s\phi\left[\left(\xi_b + \nu\xi_g\right)\right]}{EE} - 1\right]\tilde{l}^0_{b,t} \\ &\quad - \frac{s}{\nu\left(l^0_b + l^0_g\right)}\left[\frac{s\phi\xi_g}{EE} - 1\right]\tilde{l}^0_{g,t}. \end{aligned} \tag{24}$$

Because $\tilde{l}^0_{b,t}$ and $\tilde{l}^0_{g,t}$ are predetermined at time t, this equation allows us to exactly measure the on-the-job search rate $\tilde{s}_t$ consistently with the series for the job finding rate $\tilde{\phi}_t$ and the observed EE flow rate $\widetilde{EE}_t$.

With the rates $\tilde{\phi}_t$ and $\tilde{s}_t$ at hand, we can use the observed unemployment rate $\tilde{u}_{0,t}$ to pin down the fraction of bad and good matches in the next period $t+1$, using the linearized laws of motion for low- and high-quality matches in (18) and (19), which read as follows:

$$\begin{aligned}\tilde{l}^0_{b,t+1} &= -(1-\delta)\left\{\phi\xi_g l^0_b\tilde{s}_t + \left[s\xi_g l^0_b - \xi_b u_0\right]\tilde{\phi}_t\right\} \\ &\quad + (1-\delta)\left\{\left[1 - s\phi\xi_g\right]\tilde{l}^0_{b,t} + \phi\xi_b\tilde{u}_{0,t}\right\} \end{aligned} \tag{25}$$

$$\tilde{l}^0_{g,t+1} = (1-\delta)\left[\tilde{l}^0_{g,t} + \phi\xi_g l^0_b\tilde{s}_t + s\phi\xi_g\tilde{l}^0_{b,t} + \phi\xi_g\tilde{u}_{0,t} + \left[s\xi_g l^0_b + \xi_g u_0\right]\tilde{\phi}_t\right]. \tag{26}$$

With the knowledge of the distribution of match quality at time $t+1$, we can go back to equation (24) and obtain the on-the-job search rate in period $t+1$ ($\tilde{s}_{t+1}$). Repeating these steps will give us the time series of the on-the-job-search rate $\tilde{s}_t$, as well as the time series for the distribution of match quality in our sample. Note that this procedure allows us to also obtain the series of labor market slack by using equation (15).

It is important to notice that solving the model is not needed to pin down exactly the series of the on-the-job search rate. This property of the model allows us to estimate the parameters ρ_s and σ_s before solving the model. This procedure is conditioned on the fraction of bad matches at the beginning of the sample period (in our case April 1990). We assume that the distribution of match quality is at steady state at that point in time.[22]

Our approach to retrieve a model-consistent measure of the rate of the on-the-job search relies on observing the unemployment rate. We have checked that directly observing the unemployment-to-employment transition rate (UE), instead of the unemployment rate, would make it easier for the model to explain the missing inflation.[23] This is because the UE rate has improved less quickly than the unemployment rate, which implies that our model would see

[22]The results would not change if we introduced a Gaussian prior reflecting uncertainty about the initial conditions and then used the Kalman filter to optimally estimate these initial conditions.

[23]The UE flow rate is computed as in Shimer (2005).

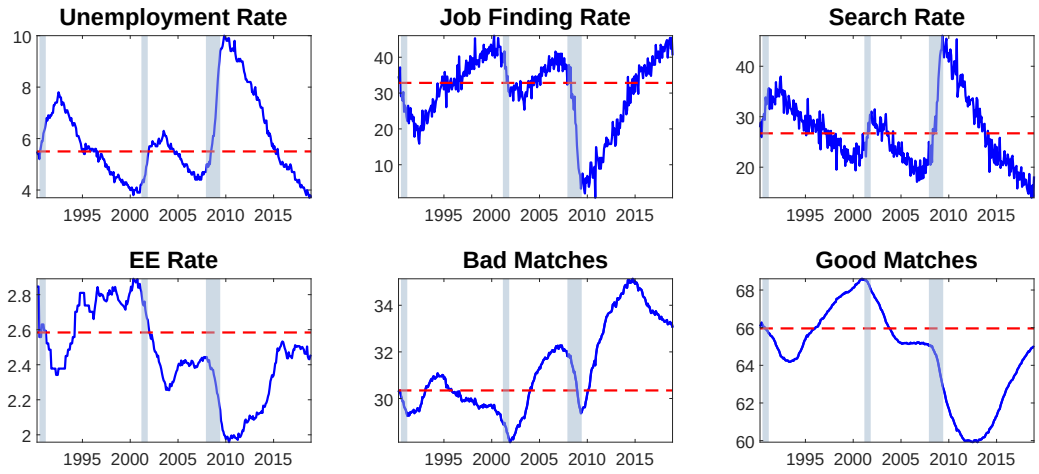

Figure 4: Labor market variables simulated from the calibrated model using the shocks that allow the model to explain exactly the observed unemployment rate and the employment-to-employment (EE) flow rate, which are shown in the left panels. The red dashed lines denote the model's steady-state value of each simulated variable. All rates are in percent.

more slack. While our measurement of the on-the-job search rate abstracts from other labor market flows, we will show that the estimated rate is remarkably close to the one measured in the micro data. This finding is very important because our mechanism and the success of our theory in explaining the missing inflation puzzle critically rest on the fall in the on-the-job search rate observed in the most recent years.

We use two monthly time series to measure the on-the-job search rate and the other labor market variables. The first series is the civilian unemployment rate measured by the U.S. Bureau of Labor Statistics (BLS). The second series is the EE flow rate measured by the *Current Population Survey* (CPS). We correct this series as suggested by Fujita, Moscarini, and Postel-Vinay (2019) and extend it back to April 1990 by splicing this series with the quit rate measured by Davis, Faberman, and Haltiwanger (2012). While the main focus of the paper is on the period that follows the Great Recession, which is when the standard theories of inflation most significantly fail, we show the behavior of the rate of on-the-job search and our measure of bad jobs over this longer period of time. We think that this exercise is interesting given that, to our knowledge, this paper is the first one that measures the search rate and the fraction of good and bad matches using aggregate labor market flows.

Figure 4 shows the dynamics of the rates of on-the-job search $\tilde{s}_t$ and bad jobs $\tilde{l}^0_{b,t}$, along with the two traditional labor market variables—the unemployment rate and the job finding rate rate $\tilde{\phi}_t$—over our sample period that goes from April 1990 through December 2018. The panels on the left report the observable variables. The traditional measures of labor market slack, such as the unemployment and the job finding rate, reported in the upper panels of Figure 4, suggest that the U.S. labor market became quite tight in recent years; however, the dynamics of two key drivers of the model's labor market slack in equation (15), i.e., the on-the job-search

rate and the stock of bad matches, paint a different picture. After the Great Recession, the rate of on-the-job search fell to a historically low level, and bad matches increased, remaining at a high level throughout the recovery. This latter finding implies that cyclical misallocation was still high, bearing down on inflation and labor productivity.

Quite interestingly, while the amount of good matches was chugging along well in recent years and was close to its long-run value, the convergence of bad matches slowed down markedly. This pattern suggests that the low unemployment rate led to the creation of a large number of low-productivity jobs that would be converted to high-productivity jobs only slowly because of the record low rate of on-the-job search.

The prediction that bad jobs were still heightened late in the sample period is consistent with the *Survey of Consumer Expectations*, which shows that about 30% of the workers employed in 2017—after eight years of recovery—were not fully satisfied with how their current job fit their experience and skills.[24] This persistent increase in bad jobs also accords well with the findings in Autor (2010), Brynjolfsson and McAfee (2011), and Jaimovich and Siu (2018), who show that job polarization intensified following the Great Recession. Jaimovich et al. (2020) document that the deterioration of employment prospects for a large share of workers who were employed in routinary occupations, led to widespread discouragement in searching for jobs. These scholars also show that one third of these workers are now working in low-paying manual nonroutinary occupations.

Using the longer sample, we find the on-the-job search rate implied by the model and reported in the top right corner of Figure 4 exhibits a clearly countercyclical pattern.[25] This countercyclicality is due to the higher volatility of the job finding rate relative to the EE rate. Because the job finding rate enters with a minus sign in the measurement equation (24) and is a strongly procyclical variable, the on-the-job search rate has to be countercyclical. Recall the fractions of good matches $\tilde{l}^0_{g,t}$ and bad matches $\tilde{l}^0_{b,t}$ are predetermined at time t and hence cannot adjust to explain the time-t changes in the job finding rate and the EE rate.

A number of explanations could support this countercyclical behavior of the on-the-job search rate. The decision to look for jobs is likely to be positively related to individual income risk, which is countercyclical. So it may be that on average, fewer employed workers search in expansions simply because less of them feel at risk of losing their jobs. To the extent that this behavior dampens the volatility of the EE rate in the data, our model rationalizes it with a countercyclical rate of on-the-job search. But the countercyclicality of on-the-job search may as well derive from compositional effects, which could also affect the dynamics of the EE flow rate in the data. Workers may search harder and hence switch jobs more often when they

[24] One of SCE questions reads as follows: "On a scale from 1 to 7, how well do you think this job fits your experience and skills?" About 30% of the respondents reported a satisfaction of 5 or less.

[25] Kudlyak and Faberman (2019) observe the job application behavior of the users of Snag-A-Job, an online job site, and find results that are consistent with the search intensity of the employed being countercyclical.

are employed in bad matches, whose number is generally big at the beginning of expansions. This view is consistent with the findings in Faberman et al. (2019), who show that employed workers search more intensively, the lower their residualized wage. In addition, workers who are hired at the beginning of an expansion are generally more skilled and dynamic than those who tend to find jobs when the labor market is already very tight. This view is consistent with the findings in Cahuc, Postel-Vinay, and Robin (2006), who show that higher-skilled workers tend to be more mobile than lower skilled ones. To the extent that these mechanisms influence the behavior of the EE flow rate in the data, the model will predict the rate of on-the-job search to be countercyclical.

One can be concerned that we do not allow the probability of drawing a good match ξ_g to vary over time. Allowing for it to do so would change our estimate of the on-the-job-search rate as this probability enters the measurement equation (23). One problem with this approach is that the large procyclicality of the job finding rate ϕ_t relatively to the EE flow rate EE_t implies that matches created in recessions are on average more productive, which is at odds with the empirical evidence reviewed by Barlevy (2002). Imposing that the probability of drawing a good match is countercyclical would reinforce our finding that the on-the-job search rate is countercyclical.

4.4 Interfirm Wage Competition and Inflation

After having characterized the in-sample dynamics of the model's labor market variables, we are in the position to validate our partial-equilibrium conjecture in Section 3.6, according to which the measure of slack Σ_t, reflecting the intensity of interfirm wage competition, is an excellent proxy for inflation in our general equilibrium model.

Equation (15) also characterizes the three components of the model's labor market slack Σ_t, which are the unemployment rate $u_{0,t}$, the measure of bad matches (capturing the degree of cyclical labor misallocation) $l^0_{b,t}$, and the share of workers searching on the job s_t. If the conjecture made in Section 3.6 is correct, then a linear combination of these three components is also key to determine the contemporaneous rate of inflation in the linearized model.

To verify this conjecture, we simulate the calibrated model for a large number of periods (one million) and then regress the simulated (annualized) series of inflation on the three determinants of slack defined in equation (15). This procedure gives us three weights that maximize the explanatory power of the three components of labor market slack on inflation. The weights are as follows:

$$\hat{\pi}_t = \underset{[-0.1719,-0.1718]}{-0.1718}\ \tilde{u}_{0,t} - \underset{[-0.0468,\ -0.0468]}{0.0468}\ \tilde{l}^0_{b,t} + \underset{[0.0419,\ 0.0419]}{0.0419}\ \tilde{s}_t, \tag{27}$$

where the numbers in square brackets under the estimated coefficients denote 95% confidence intervals.

The R-squared of the ordinary least squares (OLS) regression is 0.9993, which signifies a close-to-perfect ability of the three labor market variables to explain contemporaneous inflation in the model. This confirms our conjecture that the measure of slack in equation (15), which reflects the intensity of interfirm wage competition, is a key proxy of inflation in the model. This result is not totally surprising given the thought experiment in Section 3.6 where we studied the free-entry condition (12) in isolation, and the inspection of the impulse response functions in the previous section. While the three components in the right-hand side of equation (27) are not derived from a formal notion of output gap in the model, in practice, as we will show, they allow for a decomposition of inflation that turns out to be very useful in interpreting the results of the paper in the next section.

4.5 The Missing Inflation Puzzle Explained

We want to evaluate the ability of the model to explain the missing inflation during the recovery that followed the Great Recession. We are particularly interested in this period because the conventional theories of inflation more clearly fail to adequately account for the missing inflation, as shown in Section 2. The data set is identical to the one used to measure the on-the-job search rate in Section 4.3. We use our linearized model to retrieve the series of two shocks that make the model explain exactly the observed unemployment rate and EE flow rate. We then feed the model with these shocks to simulate the inflation rate predicted by the model in the last decade.[26]

The left panel in Figure 5 illustrates the main results of the paper by comparing the inflation rate in the data with the rate of inflation simulated from the model, as well as presenting its shock decomposition. The red line with star markers denotes the observed core PCE inflation gap, which is obtained by subtracting the ten-year PCE inflation expectations measured by the *Survey of Professional Forecasters* from the year-over-year core PCE inflation rate. The blue solid line in the left panel denotes the corresponding measure of inflation predicted by the model, using the simulation procedure described earlier. The black and white bars indicate the contributions of the shocks to the search rate and to preferences, respectively. The bars should be interpreted as the inflation rate predicted by the model when we feed it with each one of these shocks.

Unlike the traditional measures of slack analyzed in Figure 2, our alternative measure of slack reflecting the intensity of interfirm wage competition can explain the missing inflation of the past decade. The fit of the high-frequency behavior of inflation cannot be perfect because we used only two labor market variables as observables and two shocks. As illustrated by

[26] As before, we assume that the economy is in steady state at the beginning of the sample period. Different assumptions on the initial conditions would not affect our results because the sample period begins in April 1990 and the analysis focuses on a briefer period that starts several years later (specifically, in January 2011).

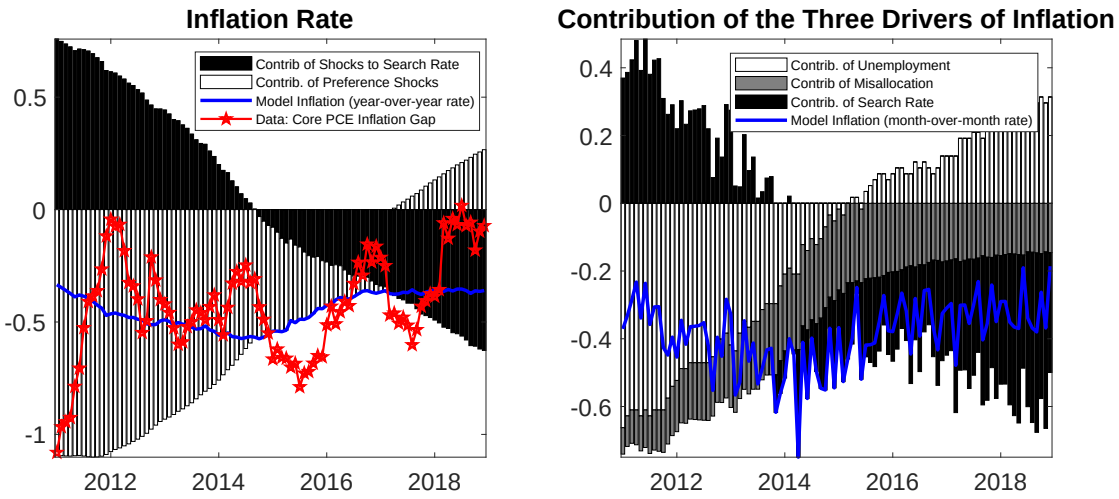

Figure 5: Left panel: core PCE inflation gap in the data (red line with the star markers) and model's corresponding inflation rate in deviation from its steady state value (blue solid line). The bars denote the contribution of the search rate shocks (black bars) and that of the preference shocks (white bars) to the model's inflation. The core PCE inflation gap is obtained by subtracting the ten-year PCE inflation expectations measured by the Survey of Professional Forecasters from the year-over-year core PCE inflation rate. The model's inflation is also computed as the year-over-year inflation rate. All rates are in percent and annualized. Right panel: Month-over-month annualized model's inflation rate (blue solid line) and the contributions of the three components of labor market slack Σ_t to this rate (bars). We use the estimated equation (27) to quantify these contributions.

the black bars, the missing inflation can be explained by the decline in the rate of on-the-job search. This drop reduced the intensity of wage competition for employed workers throughout the recovery, generating a fair amount of deflationary pressures, in spite of the steady decline in the unemployment rate. The preference shocks, which are primarily identified by the rate of unemployment, capture the state of the business cycle and the effects of the ZLB constraint. The white bars in the left panel of Figure 5 show that these factors contributed to generating deflationary pressures in the immediate aftermath of the crisis and positive inflationary pressures over the latest years of the sample when the labor market became very tight. Nevertheless, the deflationary pressures due to the decline in the rate of on-the-job search (the black bars) more than compensated for the inflationary pressures due to the preference shocks (the white bars) in latest years. In accordance with the impulse responses shown in Section 4.2, the fall in the rate of on-the-job search contributed to increasing production and to lower the rate of unemployment while exerting downward pressure on the rate of inflation.

We now use the decomposition of inflation shown in equation (27) to provide further intuition about which factors are contributing to the missing inflation. The right panel of Figure 5 visualizes the decomposition of the model's annualized month-over-month inflation rate into its three main drivers: the unemployment rate, the stock of bad jobs, and the on-the-job search rate. At the beginning of the recovery, inflation was low primarily because of the record surge in the unemployment rate during the Great Recession, as illustrated by the white bars. After 2015, further improvements in aggregate labor market conditions quickly lowered the share of unemployed job seekers, causing the unemployment rate to reverse the sign of its

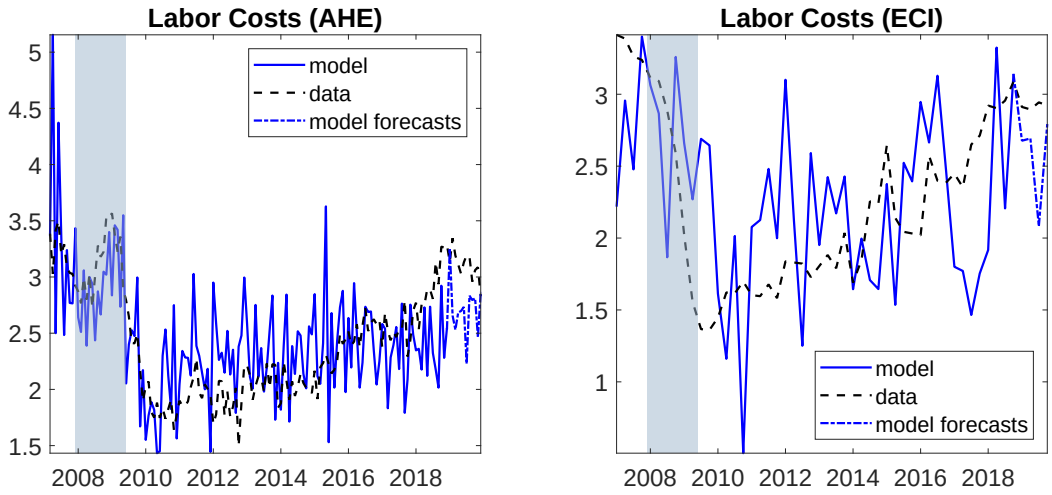

Figure 6: Left panel: Nominal marginal costs growth rate in the model and the growth rate of the average hourly earnings (AHE) of all employees: total private (CES0500000003). Moving average over the last 12 months. Source: Bureau of Labor Statistics (BLS). Units: Percentage points of annualized rates. Frequency: monthly. Right panel: Nominal marginal costs growth rate in the model and the growth rate of the Employment Cost Index (ECI): Wages and Salaries: Private Industry Workers (ECIWAG). Moving average over the last four quarters. Source: Bureau of Labor Statistics (BLS). Units: Percentage points of annualized rates. Frequency: quarterly. In both panels, the blue dotted-dashed line marks the model's forecasts of nominal marginal costs growth for the year 2019 for which we do not have data.

contribution to inflation. However, in the same years, the on-the-job search rate declined rapidly, putting downward pressures on inflation (the black bars) and dominating the effects of the unemployment rate (the white bars).

The role played by the incidence of bad matches is also very interesting (the gray bars in the right panel of Figure 5). Bad matches have always contributed to keeping inflation below its long-run level. According to the model's results, in the earlier part of the period of interest, after the unusually severe recession, a large fraction of unemployed workers took a first step onto the ladder, raising the stock of bad jobs. This development is consistent with the propagation of preference shocks to the share of bad matches shown in Figure 13 in Appendix I and is fairly typical in this class of models as it takes time, after a worker loses her job, to climb the ladder all the way up again. Later in the recovery, as the on-the-job search rate declined sharply, the speed at which workers moved to better jobs fell, exacerbating labor misallocation and keeping the intensity of wage competition low. The gray bars precisely highlight the role played by the cyclical match composition of the employment pool in explaining the missing inflation.

The Role of Labor Costs. Weak interfirm wage competition lies at the hearth of our explanation for the missing inflation puzzle. We now verify that the dynamics of U.S. wages were indeed weak in data. We do that by looking at two popular measures of U.S. workers compensation: average hourly earnings and the employment cost index from the BLS. We compare these two series with the best proxy for the effects of interfirm wage competition on price setters' costs; that is, the price of the labor service, i.e., the marginal cost. To take into

account the imperfect link between the marginal cost and the two observed measures of labor's compensation, we rescale the mean and the volatility of the model's implied growth rate of marginal costs to match those in the wage data.

The left panel of Figure 6 shows that the growth rate of nominal marginal costs predicted by the model is remarkably similar to the growth rate of average hourly earnings, even though we did not use any measure of labor costs or wages in our empirical exercise. The correlation between the two growth rates is remarkably high: 0.54. Similar conclusions can be reached by looking at the right panel of Figure 6, which compares the model's predicted growth rate of nominal marginal costs with the growth rate of the employment cost index.[27]

4.6 The On-the-Job Search Rate in the Micro Data

We now look into the micro data to validate our macro-based measurement of the on-the-job search rate. To this end, we explore a new survey that is informative about the search behavior of the employed workers; it has been administered by the Federal Reserve Bank of New York as a supplement to the *Survey of Consumer Expectations*. The SCE is a monthly and nationally representative survey of about 1,300 individuals. This survey is very useful for our purpose because it directly asks employed workers whether they have been actively searching for work in the previous seven days.[28] In this paper, we use SCE data available from 2014 trough 2017. Even if this is admittedly a short period of time, it still covers the four years in which the fall in the rate of on-the-job search predicted by our model is critical to account for the missing inflation.

Figure 7 plots the on-the-job search rate implied by the model, s_t, and the corresponding measure in the micro data (blue solid line and the black dashed-dotted line, respectively). The figure shows that the fall in the on-the-job search rate predicted by our model using aggregate labor market flows is strikingly close to the one measured in the micro data.

When the model's variable s_t is measured from equation (24), it effectively picks up a wedge between EE and UE rates, which may as well confound other effects. For instance, while the model abstracts from the intensive margin of on-the-job search, the fall in s_t measured from the macro data could potentially reflect a decline in the average number of hours spent searching. Alternatively, while the model assumes that conditional on searching, both unemployed and employed workers find jobs at the same rate $\phi(\theta_t)$, it may well be that in the data the arrival rate of job offers, conditional on searching, has diverged for these two types of job seekers, with offers becoming less frequent for the employed workers relative to the unemployed. This

[27] We show the model's forecasts for the year 2019 because our corrected series of the EE rate ends in December 2018.

[28] Question JS9 of the survey asks the following: "And within the LAST 7 DAYS, about how many TOTAL hours did you spend on job search activities? Please round up to the nearest total number of hours." We drop self-employed workers when computing the on-the-job search rate from the SCE.

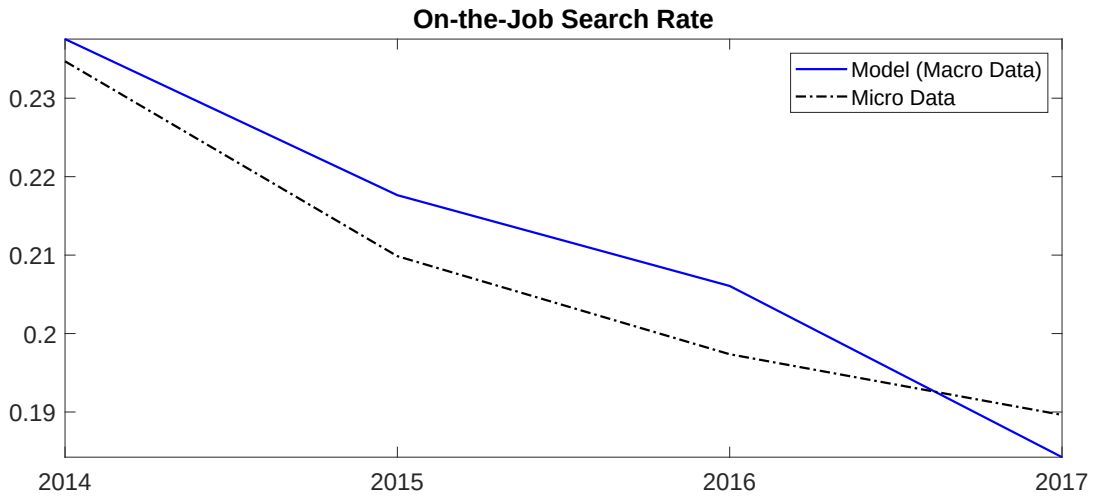

Figure 7: The on-the-job search rate in the model and in the *Survey of Consumer Expectations*. In the *Survey*, the rate is computed by dividing the workers who have searched at least one hour within the last seven days by the total number of workers surveyed. The rate is conditional on those surveyees who are working for someone else.

could be the case, for instance, if over time the employed workers had experienced a decline in the availability of suitable jobs relative to the unemployed or just faced more stringent hiring practices.

Using information on the hours of search for the employed workers in the SCE, we find that the fall in the aggregate amount of time spent searching is entirely explained by the extensive margin; that is, the effect is due to a fall in the incidence of job search among the employed—and not to a decrease in the average number of hours dedicated to search. We also looked at how the arrival rate of job offers for the employed workers varied over our sample period, relative to the arrival rate of offers for the unemployed. The evidence does not indicate a divergence in the arrival rate of offers for the employed and the unemployed.[29] Therefore, the SCE validates the decline of the on-the-job search rate predicted by our macro model over the years that are crucial to account for the lack of inflation over the post-Great Recession recovery.

5 A Model with Endogenous Search

One advantage of the model presented in Section 3 is that it allows us to derive a closed-form expression for labor market slack (Σ_t), and neatly show the role of its three driving forces: unemployment, labor misallocation, and the on-the-job search rate. An additional advantage is that labor market slack can be derived from the observation of the rates of unemployment and EE transition, without having to solve the model. A disadvantage of that model, is that it relies on the assumption of an exogenous rate of on-the-job search, where all employed worker

[29] We computed, both for the employed and the unemployed, the ratio between the total number of offers received—and not necessarily accepted—and the aggregate total number of hours spent searching.

are assumed to look for jobs with the same probability, independently of their position on the ladder and independently of the state of the business cycle. Arguably, workers want to search more actively when employed in a bad match and during expansions, when the probability of getting an offer is relatively high and its expected surplus is larger. These effects may reduce the countercyclicality of the rate of on-the-job search, which is key in our model to explain the missing inflation of the previous decade. In this section, we show that endogenizing the rate of on-the-job search does not materially affect the estimated on-the-job search rate and the distribution of match quality, leaving our conclusions on the dynamics of labor market slack and missing inflation virtually unchanged.

We construct a model that is identical to the one studied in the previous section (the benchmark model), except that agents now optimally choose whether to search on the job or not. Specifically, each period, the employed worker j draws an idiosyncratic fixed cost of search, $\varsigma_{j,t}$, from a uniform distribution

$$g\left(\varsigma_{j,t}\right) \sim \mathcal{U}\left[\xi_t \varsigma, \xi_t \varsigma + \varsigma\right], \tag{28}$$

where $\varsigma > 0$ is a parameter determining the support of the distribution and ξ_t is an aggregate shock shifting the support of this distribution. The aggregate process affecting the cost of searching on the job is assumed to follow the AR process:

$$\xi_t = \left(1 - \rho_\xi\right) \bar{\xi} + \rho_\xi \xi_{t-1} + \varepsilon_{\xi,t}, \quad \varepsilon_{\xi,t} \backsim \boldsymbol{N}\left(0, \sigma_\xi\right), \tag{29}$$

where $\bar{\xi}$ is parameter capturing the unconditional mean of the process ξ_t. A negative realization of this shock shifts the support of the distribution downward, raising the probability of drawing a fixed cost of search that is lower than the expected return. Consequently, more agents would search on the job. While the shock, ε_ξ, affects the distribution of the fixed cost $\varsigma_{j,t}$ identically across workers, the individual response of a worker's propensity to search on the job depends on their position on the ladder and on the share of surplus they are able to extract from their current match.

The worker-specific costs $\varsigma_{j,t}$ are purely psychological and do not absorb households' resources. It should be noted that the lower bound of the support of the uniform distribution in equation (28) can take negative values, implying that employed workers may search even if they do not expect an increase in surplus. A negative shock should be interpreted as a psychological reward to changing jobs that is unrelated to its compensation and is unobservable by the firm. The aggregate shock ξ_t is necessary to redo the empirical analysis of Section 4.5, which requires the model to have at least two shocks to jointly explain the two labor-market variables (the unemployment rate and the EE rate). The other shock is the shock to households' preferences.

In this economy, every worker is willing to search on the job, provided that the stochastic

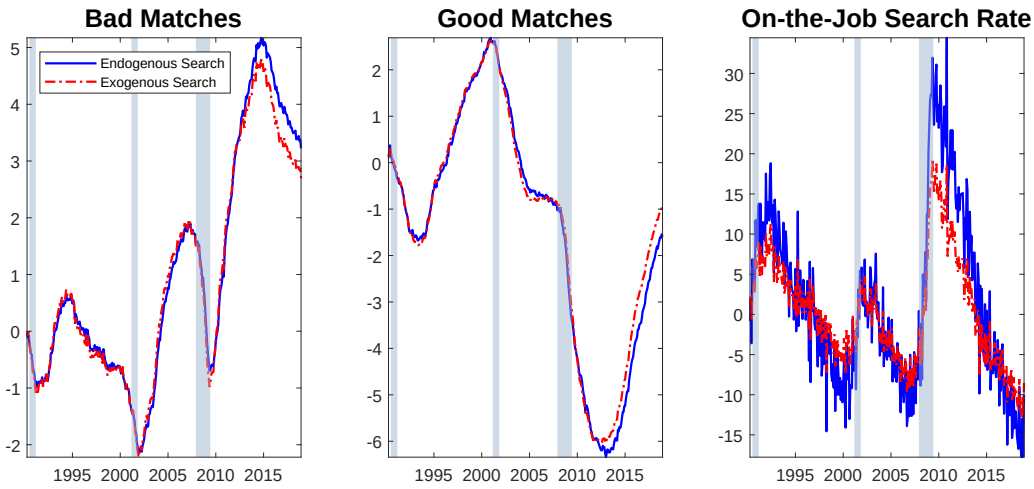

Figure 8: The in-sample dynamics of bad matches (left plot), good matches (central plot), and the on-the-job search rate (right plot) according to the benchmark model with exogenous on-the-job search (red dashed-dotted line) and the model with endogenous on-the-job search (blue solid line). Units: percentage points in deviation from their steady-state value.

cost of search, $\varsigma_{j,t}$, is below the expected return, i.e. the expected gain in surplus. The return to search will depend on (i) the position of workers on the ladder, (ii) the amount of match surplus that they are able to extract, and (iii) the state of the business cycle, which affects both the probability of meeting a poaching firm and the expected surplus from being poached. The bargaining protocol − based on Bertrand wage competition among firms − is the same as in the benchmark model. In this setup, we need to track the decision to search for five types of workers: (i) workers who are employed in a bad match and receive no surplus; (ii) workers who are in a good match and receive no surplus; (iii) workers who are in a bad match under full surplus extraction; (iv) workers who are in a good match with partial extraction of surplus (i.e., they get the surplus of a bad match); (v) workers who are employed with full extraction of a good-match surplus.[30] For each of the five categories of workers listed above, the fraction of workers who search on the job is determined by the uniform cumulative distribution; in symbols, $0 <\text{Prob}\{\varsigma_{j,t} < E_t\Delta S_t(k)\} = \frac{E_t\Delta S_t(k)}{\varsigma} - \xi_t < 1$, where $E_t\Delta S_t(k)$ denotes the expected surplus gain from on-the-job search for each of the five types of workers ($k = \{i, ii, iii, iv, v\}$). More details about the model are provided in Appendix J.

In this extended model, a positive preference shock, by increasing the probability of meeting a poaching firm and hence the expected return from on-the-job-search, increases the fraction of employed workers who look for jobs. This is a key departure from the benchmark model, in which the on-the-job search rate is exogenous and orthogonal to preference shocks. These effects could potentially make the on-the-job search rate less countercyclical and, consequently, may impair the ability of our theory to explain the missing inflation.

[30] Assuming that workers who are extracting the full surplus of their good match do not search would not materially affect the results of this robustness exercise.

To calibrate the model, we follow the same approach discussed in Section 4.1. This implies that all parameters that are common to the model with exogenous search will take the values reported in Table 1. There are only four parameters in the model with endogenous search that do not feature in the baseline model, and therefore need to be discussed here; they pertain to equations (28) and (29) above. The persistence and the standard deviation of the shocks to the distribution of the cost of on-the-job search, ξ_t, is estimated using a least-square approach.[31] The estimation leads to $\rho_\xi = 0.7524$ and $100\sigma_\xi = 5.08$. The parameters $\bar{\xi}$ and ς are set to -0.1452 and 0.5583, respectively, so as to ensure that the steady-state rate of EE transitions and the fraction of employed job seekers are consistent with the corresponding values in the model with exogenous on-the-job search.

Figure 8 compares the dynamics of bad matches, good matches, and the rate of on-the-job search (in percentage deviations from their steady-state value) obtained in the model with exogenous and endogenous on-the-job search (red dashed-dotted and blue solid lines, respectively). The two models deliver very similar results, with the endogenous-search model suggesting a somewhat more countercyclical on-the-job search rate and a somewhat larger slack at the end of the sample, as reflected in a larger mass of bad matches and a lower rate of on-the-job search rate. Similarly to the benchmark model, the model with endogenous search accounts for the missing inflation of the last decade.

The reason why the two models provide similar predictions lies in the lower procyclicality of the EE rate relative to the UE rate. This feature of the data is explained in both models by the countercyclical search behavior of the employed. In the model with endogenous on-the-job search, the stagnant EE flow during the past economic recovery is largely accounted for by the decline in the search rate of the workers employed in bad matches under full surplus extraction; these are workers who remained stuck in their bad match, unable to generate outside offers from good-match firms. Their low propensity to search sustained the creation of a large mass of low-productivity jobs, which had the effect of lowering further the intensity of interfirm wage competition (see Appendix K). This prediction is consistent with the hypothesis of discouragement, due to the growing struggle of a large share of American workers to find the type of job they had before the Great Recession (Jaimovich et al., 2020).

[31] We make an initial guess on the parameters ρ_ξ and σ_ξ and then use the Kalman filter to obtain the estimated series for the stochastic process of the shocks $\tilde{\xi}_t \equiv \left(\xi_t - \bar{\xi}\right)$ implied by the calibrated model and the data (the unemployment rate and the EE flow rate). We then regress the estimated series of $\tilde{\xi}_t$ on $\tilde{\xi}_{t-1}$ and obtain the OLS estimate of the parameters ρ_ξ and σ_ξ, which are then used to check our guess. If these values diverge, we update our guess using the last OLS estimate of these parameters and redo the procedure until convergence.

6 Empirical Performance in a Longer Sample Period

In this section, we discuss how well our measure of slack, defined in equation (15), fares at explaining inflation dynamics relative to other popular measures when the sample period is extended to include the 1990s and the 2000s. The availability of data on EE flows prevents us to extend our analysis to earlier decades.

We find that the ability of our measure of slack to explain inflation dynamics is better than the traditional ones in this longer sample period. We do so in the simplest possible way, which is to compare how our measure of slack performs relative to other traditional theory-based ones, using standard Phillips Curve regressions. That is, we estimate the equation

$$\pi_t = \beta \cdot slack_t + \varepsilon_t, \tag{30}$$

where π_t is the eight-quarter moving average of the quarter-over-quarter core PCE inflation rate (annualized and in percent) and in deviation from 2%, which is assumed to be the long-run value for core PCE inflation.[32] We use the moving average as we are not interested in fitting the high-frequency swings in inflation. The variable $slack_t$ represents different measures of labor market slack: our own Σ_t, based on the intensity of wage competition, and each of the measures considered in Section 2—that is, the labor share; a version of the labor share augmented to account for search and matching frictions; the unemployment gap; and detrended total hours, which is the key observable to inform the output gap in state-of-the-art DSGE models, such as those in Christiano, Eichenbaum, and Evans (2005), Smets and Wouters (2007), and Justiniano, Primiceri, and Tambalotti (2010). After having estimated the Phillips curve (30) for the period 1990Q2 through 2018Q4, we compute the root mean squared error (RMSE) of the different specifications over different subsamples.

In this longer sample period, our measure of slack attains the smallest RMSE. The particularly appalling performance of the alternative measures of slack in the last decade are the main driver of this result, justifying our emphasis on the post-Great Recession recovery in the structural analysis of this paper.

7 Concluding Remarks

We showed that standard theories of inflation based on the New Keynesian Phillips curve fail to explain why inflation has remained subdued throughout the post-Great Recession recovery. We introduced a model with the job ladder in which the fraction of workers searching on the

[32] We cannot use the *Survey of Professional Forecasters*' expectations of PCE inflation over the next ten years to compute the inflation gap as we did in our empirical analysis that focused on the last decade. The reason is that this measure of long-term inflation expectations became available only since 2008.

job influences labor market slack by affecting the degree of interfirm wage competition to hire employed workers. We found that the model explains the missing inflation of the past decade with the fall in the rate of on-the-job search and the associated weakening of wage competition among firms. Finally, we verified that when the on-the-job search rate is identified at micro levels using survey data, a similar fall in this rate is detected from 2014 through 2017.

Our paper opens avenues for future research on the appropriate stabilization policies in the presence of interfirm competition for the employed. For instance, an important question to explore is whether monetary policy, whose primary goal is to stabilize inflation, has any significant effect on the search behavior of the employed. While the empirical literature has made important progress in understanding how monetary impulses affect labor supply mobility, very little is known about the effectiveness of monetary stimuli in incentivizing workers to search on the job.

Our contribution is to show that accounting for on-the-job search improves our understanding of inflation dynamics. Dissecting among its various potential determinants is an active area of empirical labor research. Our finding implies that making further progress in this direction is all the more important, as it contributes to our understanding of macroeconomic dynamics, beyond the specificity of the labor market.

References

Abraham, G. K., and C. J. Haltiwanger (2019): "How Tight is the Labor Market?," Federal reserve bank of chicago working paper.

Atkeson, A., and L. E. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?," *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), 2–11.

Autor, D. (2010): "The Polarization of Job Opportunities in the U.S. Labor Market: Implications for Employment and Earnings," Report, Brookings Institute, April 30 2010.

Barlevy, G. (2002): "The Sullying Effect of Recessions," *Review of Economic Studies*, 69(1), 65–96.

Brynjolfsson, E., and A. McAfee (2011): *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press.

Cahuc, P., F. Postel-Vinay, and J.-M. Robin (2006): "Wage Bargaining with On-the-Job Search: Theory and Evidence," *Econometrica*, 74(2), 323–364.

Christiano, L., M. Eichenbaum, and C. Evans (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

Christiano, L. J., M. S. Eichenbaum, and M. Trabandt (2016): "Unemployment and Business Cycles," *Econometrica*, 84, 1523–1569.

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

Del Negro, M., M. Lenza, G. Primiceri, and A. Tambalotti (2020): "What's up with the Phillips Curve," *Brookings Papers on Economic Activity*, Spring, 301–357.

Dotsey, M., and R. G. King (2005): "Implications of state-dependent pricing for dynamic macroeconomic models," *Journal of Monetary Economics*, 52(1), 213–242.

Faberman, R. J., A. I. Mueller, A. Sahin, and G. Topa (2017): "Job Search Behavior among the Employed and Non-Employed," IZA Discussion Papers 10960, Institute of Labor Economics (IZA).

Fernald, J. G. (2016): "Reassessing Longer-Run U.S. Growth: How Low?," Working Paper 2016-18, Federal Reserve Bank of San Francisco.

Fujita, S., G. Moscarini, and F. Postel-Vinay (2019): "Measuring Employer-to-Employer Reallocation," Working paper.

Galí, J., and M. Gertler (1999): "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, pp. 195–222.

Galí, J., F. Smets, and R. Wouters (2011): "Unemployment in an Estimated New Keynesian Model," in *NBER Macroeconomics Annual 2011, Volume 26*, NBER Chapters, pp. 329–360. National Bureau of Economic Research, Inc.

Gertler, M., C. Huckfeldt, and A. Trigari (2019): "Unemployment Fluctuations, Match Quality, and the Wage Cyclicality of New Hires," Review of Economic Studies, forthcoming.

Gertler, M., L. Sala, and A. Trigari (2008): "An Estimated Monetary DSGE Model with Unemployment and Staggered Nominal Wage Bargaining," *Journal of Money, Credit and Banking*, 40(8), 1713–1764.

Jaimovich, N., I. Saporta-Eksten, H. Siu, and Y. Yedid-Levi (2020): "The Macroeconomics of Automation: Data, Theory, and Policy Analysis," mimeo.

Jaimovich, N., and H. Siu (2018): "The Trend is the Cycle: Job Polarization and Jobless Recoveries," Review of Economics and Statistics, forthcoming.

Justiniano, A., G. E. Primiceri, and A. Tambalotti (2010): "Investment shocks and business cycles," *Journal of Monetary Economics*, 57(2), 132–145.

Kimball, M. S. (1995): "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 27(4), 1241–1277.

Krause, M. U., D. Lopez-Salido, and T. A. Lubik (2008): "Inflation dynamics with search frictions: A structural econometric analysis," *Journal of Monetary Economics*, 55(5), 892–916.

Kudlyak, M., and R. J. Faberman (2019): "The Intensity of Job Search and Search Duration," *American Economic Journal: Macroeconomics*, 11(3), 327–57.

Levin, A., J. D. Lopez-Salido, and T. Yun (2007): "Strategic Complementarities and Optimal Monetary Policy," CEPR Discussion Papers 6423, C.E.P.R. Discussion Papers.

Lindé, J., and M. Trabandt (2018): "Resolving the Missing Deflation Puzzle," mimeo, Freie Universitat Berlin.

McLeay, M., and S. Tenreyro (2019): "Optimal Inflation and the Identification of the Phillips Curve," in *NBER Macroeconomics Annual 2019, volume 34*. National Bureau of Economic Research, Inc.

Moscarini, G., and F. Postel-Vinay (2016): "Did the Job Ladder Fail after the Great Recession?," *Journal of Labor Economics*, 34(S1), 55–93.

——— (2018): "On the Job Search and Business Cycles," Mimeo Yale U. and U. College London.

——— (2019): "The Job Ladder: Inflation vs Reallocation," Mimeo Yale U. and U. College London.

Phillips, W. (1958): "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861 - 1957," *Economica*, 25(100), 293–299.

Postel-Vinay, F., and J.-M. Robin (2002): "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity," *Econometrica*, 70(6), 2295–2350.

Ravenna, F., and C. E. Walsh (2008): "Vacancies, unemployment, and the Phillips curve," *European Economic Review*, 52(8), 1494–1521.

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

Silva, J. I., and M. Toledo (2009): "Labor Turnover Costs and the Cyclical Behavior of Vacancies and Unemployment," *Macroeconomic Dynamics*, 13(S1), 76–96.

Sims, C., and T. Zha (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39(4), 949–68.

Smets, F., and R. Wouters (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *The American Economic Review*, 97(3), 586–606.

Stock, J. H., and M. W. Watson (2007): "Why Has U.S. Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking*, 39(s1), 3–33.

Stock, J. H., and M. W. Watson (2008): "Phillips Curve Inflation Forecasts," Working Paper 14322, National Bureau of Economic Research.

——— (2019): "Slack and Cyclically Sensitive Inflation," Working Paper 25987, National Bureau of Economic Research.

Appendices (Not For Publication)

In Appendix A, we show the acceptance ratio in the data. We summarize how to construct the measure of marginal costs in a standard New Keynesian model in Appendix B. Different calibrations and specifications for the Phillips curve studied in Section 2 of the main text are introduced, and their ability to account for the missing inflation after the Great Recession is evaluated in Appendix C, which focuses on Phillips curves with a backward-looking component. In Appendix D we describe how the data set to conduct the VAR analysis in Section 2 of the main text is constructed. In Appendix E, we show that state-of-the-art dynamic general equilibrium models have hard time explaining the missing inflation. We show how to work out equations (13) and (14) in the main text, which provide an analytical characterization of the surpluses in the model, in Appendix F. In Appendix G, we show the robustness of our main results by varying two parameters that are hard to calibrate: the probability of meeting a worker that is a bad match for the firm (ξ_b) and the probability that workers switch jobs if they receive an outside offer that makes them indifferent (ν). We also show how the results change when varying the persistence of the on-the-job search rate (ρ_s). In Appendix H, we show how we solve the model with an occasionally binding zero lower bound for the nominal interest rate. In Appendix I we show how preference shocks propagate. Finally, we provide more specifics about the model in which agents optimally decide whether to search on the job and its predictions in Appendices J and K.

A Acceptance Ratio

Figure 9 shows the ratio of the employment-to-employment flow rate, corrected as suggested by Fujita, Moscarini, and Postel-Vinay (2019), to the unemployment-to-employment flow rate.[33] This plot shows that the acceptance ratio rapidly rose during the Great Recession. However, the acceptance ratio steadily decreased during the recovery and eventually moved below its pre-Great Recession average computed over the period from February 1996 through December 2007, which is denoted by the red dashed line.[34]

Moscarini and Postel-Vinay (2019) interpret this ratio as the acceptance ratio. In their model, because the fraction of accepted offers is higher when more workers are employed in low-productivity jobs, this ratio is a proxy for the degree of labor misallocation and is inversely related to inflation in their model. When this ratio is low, few offers are accepted on average as labor is perfectly allocated and, as result, marginal costs and inflation are high in their model.

[33] The correction proposed by Fujita, Moscarini, and Postel-Vinay (2019) ends up revising the employment-to-employment rate upward in recent years, causing the fall of this ratio to be less rapid and dramatic during the post-Great Recession recovery than one would obtain by using the uncorrected CPS series for the EE flow rate.

[34] The CPS data start in February 1996.

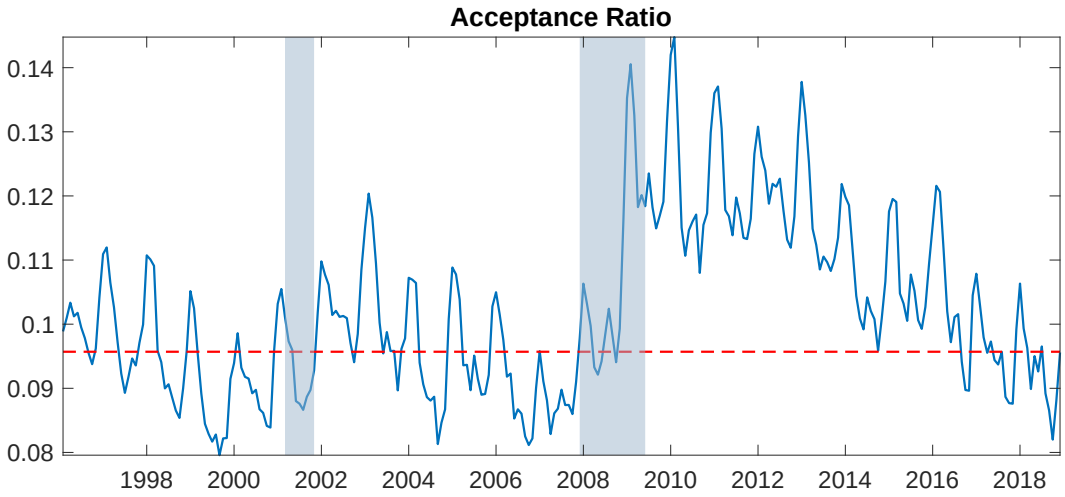

Figure 9: Acceptance Ratio. The ratio of the employment-to-employment flow rate to the unemployment-to-employment flow rate. Both rates are computed by taking the three-month moving average of the Current Population Survey (CPS) flow data. The red dashed line denotes the mean of the ratio computed from February 1996 through December 2007. The employment-to-employment rate is corrected as proposed by Fujita, Moscarini, and Postel-Vinay (2019).

In our model, a low acceptance ratio may be due to either a high degree of misallocation or a low share of workers searching on the job. Therefore, this ratio is not always a good predictor of labor misallocation and inflation in our model. A better predictor is the empirical measure of labor market slack, which is based on the intensity of interfirm wage competition, introduced in Section 3.6.

B Computation of Real Marginal Costs in a Standard New Keynesian Model with Search and Matching

We follow the work by Krause, Lopez-Salido, and Lubik (2008), who study the behavior of real marginal costs in a simple New Keynesian model with search and matching frictions in the labor market. Equation (32) from Krause, Lopez-Salido, and Lubik (2008, p. 898) defines the real marginal cost as:

$$mc_t = \frac{W_t}{\alpha\left(\frac{y_t}{n_t}\right)} + \frac{c'\left(v_t\right)/q\left(\theta_t\right) - \left(1-\rho\right)E_t\beta_{t+1}c'\left(v_{t+1}\right)/q\left(\theta_{t+1}\right)}{\alpha\left(\frac{y_t}{n_t}\right)}, \tag{31}$$

where W_t denotes the real hourly wage, y_t/n_t is the average product of labor, $c'\left(v_t\right)$ is the derivative of the vacancy cost function with respect to vacancies, $q\left(\theta_t\right)$ is the vacancy filling rate, β_{t+1} is the discount factor, and α is the elasticity of output to employment in the production function. The first component on the right-hand side of equation (31) is the unit labor cost, i.e., the ratio of the labor cost and the marginal product of labor. The second component stems

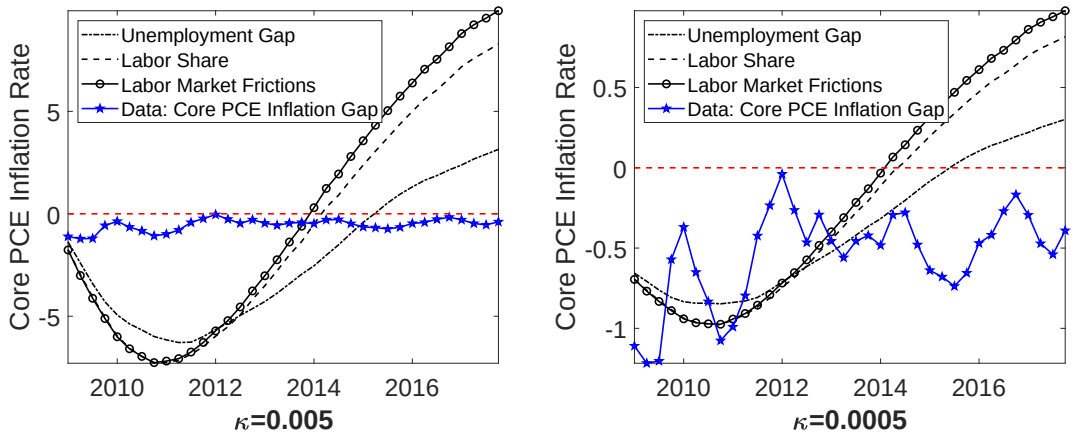

Figure 10: PCE refers to the Price Index for Personal Consumption Expenditures. The figure shows the inflation dynamics from 2009Q1 through 2017Q3 using three traditional theories of inflation using a slope of the inertial Phillips curve of 0.005 (left panel) and 0.0005 (right panel) and the core PCE inflation gap. The assumed degree of inertia is equal to 0.90.

from the existence of search and matching frictions and can be interpreted as cost savings from not having to hire in the following period.

Let $s_t \equiv W_t/\alpha \left(\frac{y_t}{n_t}\right)$ denote the unit labor cost, which equals the labor share of income divided by the elasticity of output to employment. Krause, Lopez-Salido, and Lubik (2008) show that linearizing equation (31) and rearranging leads to the following expression:

$$\widehat{mc_t} = \hat{s}_t + \frac{1-\phi}{1-\tilde{\beta}}\left[\frac{\xi}{1-\xi}\left(\hat{h}_t - \tilde{\beta}E_t\hat{h}_{t+1}\right) + (\varepsilon_c - 1)\left(\hat{v}_t - \tilde{\beta}E_t\hat{v}_{t+1}\right) - \tilde{\beta}E_t\hat{\beta}_{t+1} - \left(1-\tilde{\beta}\right)\hat{w}_t\right], \tag{32}$$

where a hat variable is used to denote log deviations from the steady-state, h_t denotes the job finding rate, $\tilde{\beta}$ is a discount factor adjusted for the rate of job separation, ε_c is the elasticity of vacancy costs to vacancies, ξ is the elasticity of the matching function with respect to unemployment, and $\phi = s/mc$ is the share of unit labor cost over total marginal costs. We follow the calibration in Krause, Lopez-Salido, and Lubik (2008) and assume that $\xi = 0.5$, $1-\phi = 0.05$, and $\tilde{\beta} = 0.943$. In line with the model specified in Section 3, we assume a linear vacancy cost function, which implies $\varepsilon_c = 1$, and log utility in consumption.

C Traditional Measure of Slack: Robustness

The most popular Phillips curve used in empirical studies features a backward-looking term:

$$\pi_t = \iota\pi_{t-1} + \kappa\varphi_t + E\pi_{t+1}, \tag{33}$$

where the parameter ι reflects the degree of price indexation. We redo the same VAR-based exercise as the one in Section 2 to evaluate the robustness of these results to the introduction

of price indexation and of a flatter Phillips curve. We kick off by setting the degree of price indexation ι_p to 0.9—an upper bound for plausible degrees of inertia. We consider two cases. In the first case, we assume that the slope of the Phillips curve is $\kappa = 0.005$, as in the baseline case analyzed in the main text. In the second case, we consider a very flat Phillips curve with a parameter $\kappa = 0.0005$. While the first case allows us to evaluate how much adding a backward-looking component alters the result shown in the main text (Figure 2), the second case is useful to illustrate that even a very flat Phillips curve with a lot of price indexation cannot solve the missing inflation puzzle.

The results are shown in Figure 10. In the left panel we show the first case, which confirms that adding price indexation just makes the drop in inflation in 2009 more pronounced and delayed. According to the traditional measures of slack, inflation should have picked up by 2014; only the unemployment gap predicts inflation staying below its long-run level until the end of 2015. None of the measures explains why inflation has not risen after nine years of economic expansion, even after introducing a very large degree of price indexation.

In the right panel of Figure 10 we show the second case, which combines a large degree of price indexation with an extremely flat slope of the Phillips curve. Comparing the plots of the left and right panels of Figure 10 reveals that a reduction of the slope in the presence of high indexation decreases the predicted fall in inflation at the beginning of the sample and contains the predicted rise in inflation at the end of the sample. Therefore, we conclude that a flatter Phillips curve in and of itself does not solve the puzzle of the persistently low inflation observed in the past decade. This is the case even if one endows the Phillips curve with a very large price inertia.

D Construction of the Time Series and Their Sources

The time series used for the VAR analysis have been constructed from the following data downloaded from the St. Louis Fed's database called Federal Reserve Economic Data (FRED). The labor share of income is computed as the ratio of total compensation in the nonfarm business sector divided by nominal nonfarm GDP. In turn, total compensation is computed as the product of compensation per hour (COMPNFB) times total hours (HOANBS), and nominal GDP is the product of real output (OUTNFB) times the appropriate deflator (IPDNBS). All series are quarterly and seasonally adjusted. We compute the deviations of the labor share from its trend by computing log deviations from an eight-year moving average.

We follow Shimer (2005) and compute the job finding rate as $\phi_t = 1 - \left(u_{t+1} - u^s_{t+1}\right)/u_t$, where u^s_{t+1} denotes the number of workers employed for less than five weeks in month $t+1$ (UEMPLT5). The total number of workers unemployed in each month is computed as the sum of the number of civilians unemployed less than five weeks (UEMPLT5), for 5 to 14

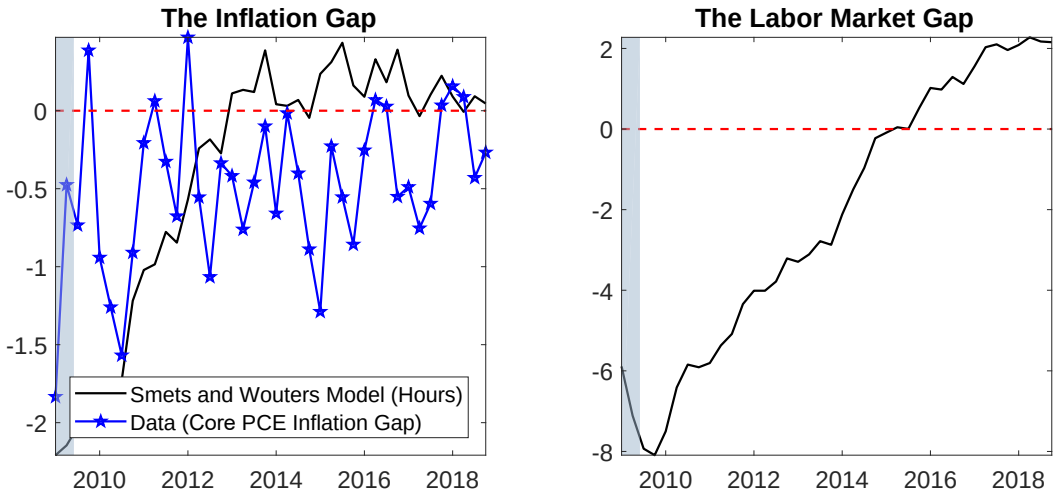

Figure 11: Left panel: PCE refers to the Price Index for Personal Consumption Expenditures. The black solid line: Inflation predicted by the model of Smets and Wouters (2007) conditional on observing the series of hours plotted on the right panel. The blue starred line: the difference between the annualized quarter-to-quarter core PCE inflation rate and the ten-year-ahead core PCE inflation expectations based on the Survey of Professional Forecasters. Right panel: Hours worked detrended using their eight-year moving average.

weeks (UEMP5TO14), 15 to 26 weeks (UEMP15T26), and 27 weeks and over (UEMP27OV). The primary data are constructed by the U.S. Bureau of Labor Statistics from the CPS and seasonally adjusted. To obtain quarterly percentage point deviations of the job finding rate from its trend, we average monthly data over each quarter and then subtract the actual job finding rate from its eight-year moving average.

We also use data on real gross domestic product (GDPC1), real gross private domestic investment (GDPIC1), and real personal consumption expenditures (PCECC96). All data are quarterly and seasonally adjusted. When computing percentage deviations of each of these times series from its trend, we first remove a quadratic trend from the variable in logs and then take the difference from its eight-year moving averages. To compute percentage deviations of real wages from the trend we first remove a linear trend to the log of compensation per hour (COMPNFB) and then take the difference with respect to its eight-year moving average.

We measure aggregate price inflation by taking log differences on the previous quarter of the seasonally adjusted Consumer Price Index for All Urban Consumers (CPIAUCSL). We also use quarterly data on the effective federal funds rate (FFR) and on the short-term natural rate of unemployment (NROUST). We compute percentage point deviations of inflation, the federal funds rate, and the natural rate of unemployment from trend as the difference from each series' eight-year moving average.

E A State-of-the-Art Dynamic General Equilibrium Model (Smets and Wouters 2007)

In this appendix, we evaluate the ability of a leading empirical general equilibrium model to reconcile labor market and inflation dynamics in the post-Great Recession recovery. We use the popular model introduced by Smets and Wouters (2007) to perform this exercise. This is a model with many real and nominal frictions and a large array of shocks and is well known to fit the U.S. macro series well. Smets and Wouters conduct a Bayesian estimation of the parameters of their model using seven observables: consumption growth, investment growth, GDP growth, hours (detrended for the labor force participation), inflation, real wage, and the federal funds rate. Their sample period goes from 1966Q1 through 2004Q4. We extend their data set trough 2018Q4 and detrend the series of hours using a eight-year moving average. We make the latter change because the series of hours exhibited a significant downward shift since the onset of the Great Recession and has never attained its pre-recession level again.

We use the extended data set to estimate the model. Then the same data set is used to filter the state variables of the estimated model from the first quarter of 1966 through the fourth quarter of 2008. For the subsequent periods (2009Q1–2018Q4), we filter the state variables of the estimated model using only the series of hours in order to obtain inflation predictions conditional on labor market data only. Recall that the emphasis of this paper is on the apparently waning link between the labor market and inflation. The black solid line in the right panel of Figure 11 shows the series of hours detrended using a eight-year moving average, which we use to simulate the Smets and Wouters model.

Based on the series of hours, the Smets and Wouters' model predicts that inflation is above target already in 2012. See the black solid line in the left panel of Figure 11. The plot also reports the inflation gap in the data (blue starred line), which is computed by taking the difference between the annualized quarter-to-quarter core PCE inflation rate and the ten-year-ahead core PCE inflation expectations based on the *Survey of Professional Forecasters*. The inflation gap in the data remains persistently below zero, whereas the Smets and Wouters' model predicts that inflation moves above its long-run level as early as in 2012. Indeed, the right panel of Figure 11 shows that the series of hours implied that the labor market became tight (positive labor market gap) in 2015.

F Job Values and Sequential Auctions

In this section we derive the expressions for the surplus function $S_t(y)$ in equation (13), following the approach in Moscarini and Postel-Vinay (2019). We start by characterizing the value functions for the states of employment and unemployment. The value of unemployment to a

worker j measured after worker reallocation has taken place and expressed in utility units is determined as follows::

$$\lambda_t V_{u,t}^j = b + \beta E_t \phi\left(\theta_{t+1}\right) \lambda_{t+1} \left[V_{e,t+1}^j \left(w_{t+1}\left(j\right), y_{t+1}\left(j\right) \mid e_{t+1}^0 = 0\right)\right] + \beta E_t \left(1 - \phi\left(\theta_{t+1}\right)\right) \lambda_{t+1} V_{u,t+1}^j, \tag{34}$$

where we let the indicator function $e_{t+1}^0 = \{0, 1\}$ denote the state of employment at the beginning of period $t+1$, before reallocation takes place.

The value to a worker j of being employed at the production stage of period t in a job of productivity y_t at wage w_t after reallocation has taken place, but before the realization of the current-period separation shock, is determined as follows:

$$\begin{aligned}
\lambda_t V_{e,t}^j \left(w_t\left(j\right), y_t\left(j\right)\right) &= \lambda_t \frac{w_t\left(j\right)}{P_t} + \beta E_t \lambda_{t+1} \left\{\delta\left[1 - \phi\left(\theta_{t+1}\right)\right] V_{u,t+1}^j \right. \\
&+ \delta \phi\left(\theta_{t+1}\right) V_{e,t+1}^j \left(w_{t+1}\left(j\right), y_{t+1}\left(j\right) \mid e_{t+1}^0 = 0\right) \\
&\left. + \left(1 - \delta\right) V_{e,t+1}^j \left(w_{t+1}\left(j\right), y_{t+1}\left(j\right) \mid w_t\left(j\right), y_t\left(j\right), e_{t+1}^0 = 1\right)\right\}.
\end{aligned} \tag{35}$$

The above expression implies that the worker receives a wage $\frac{w_t(j)}{P_t}$ in exchange for her labor services, plus a continuation value, which depends on whether the worker separates or not at the end of the period. If separation occurs at rate δ, the worker will still be in the state of unemployment by the end of period $t+1$ if no job is found, which occurs with probability $1 - \phi\left(\theta_{t+1}\right)$. In this case the worker receives the expected present value $E_t V_{u,t+1}^j$. If instead the newly separated worker finds a job in period $t+1$ with probability $\phi\left(\theta_{t+1}\right)$, she gets the payoff of being in a match of productivity $y_{t+1}\left(j\right)$, paying the wage $w_{t+1}\left(j\right)$, which is conditional on the worker having separated at the end of time t and therefore being unemployed at the beginning of $t+1$. The expected present discounted value of a such job, expressed in units of the numeraire good is denoted by $E_t V_{e,t+1}\left[w_{t+1}\left(j\right), y_{t+1}\left(j\right) \mid e_{t+1}^0 = 0\right]$. With probability $1 - \delta$ instead, the worker does not separate at the end of time t, receiving $E_t V_{e,t+1}\left[w_{t+1}\left(j\right), y_{t+1}\left(j\right) \mid w_t\left(j\right), y_t\left(j\right), e_{t+1}^0 = 1\right]$ at the end of the next period. This expression captures the value of being employed at the end of time $t+1$ in a match with productivity y_{t+1} at the wage w_{t+1}, conditional on having been employed in a match with productivity $y_t\left(j\right)$ and wage $w_t\left(j\right)$ in the previous period and not having separated between periods, i.e., being in employment at the beginning of period $t+1$. Note that this expected value includes the possibility of a job-to-job transition in period $t+1$.

We assume that firms have all the bargaining power, and hence, the unemployed workers who take up a new offer are indifferent between being employed or unemployed, i.e.,

$$\lambda_t V_{e,t}\left(w_t\left(j\right), y_t\left(j\right) \mid e_t^0 = 0\right) = b + \beta E_t \lambda_{t+1} V_{u,t+1}^j \tag{36}$$

independently of $y_t(j)$. It follows that

$$V_{u,t}^j = \frac{b}{\lambda_t} + \beta E_t \frac{\lambda_{t+1}}{\lambda_t} V_{u,t+1}^j = V_{u,t}. \tag{37}$$

Let $V_{e,t}^*(y)$ denote the value to the worker of being employed under full extraction of a firm's willingness to pay at the end of time t. In this case a worker of productivity y receives the maximum value that the firm is willing to promise in period t, including the payment of the current-period wage. Let $\{w_s^*(y)\}_{s=t}^{\infty}$ denote the state-contingent contract that delivers $V_{e,t}^*(y) \equiv V_{e,t}(w_t^*, y)$. By promising to pay the contract $\{w_s^*(y)\}_{s=t}^{\infty}$, the firm breaks even in expectation, that is, the expected present value of future profits is zero.

Now consider a firm that is currently employing a worker with productivity y under any promised contract $\{w_s(y)\}_{s=t}^{\infty}$. Assume that the worker is poached by a firm with match productivity y'. The outcome of the auction must be one of the following three:

1. $V_{e,t}^*(y') < V_{e,t}(w_t, y)$; in this case the willingness to pay of the poaching firm is less than the value of the contract that the worker is currently receiving. As a result, the incumbent firm retains the worker with the same wage contract with value $V_{e,t}(w_t, y)$.

2. $V_{e,t}(w_t, y) \leq V_{e,t}^*(y') < V_{e.t}^*(y)$; in this case the willingness to pay of the poaching firm is greater or equal to the value of the contract the worker is receiving in his current job, but lower than the willingness to pay of the incumbent firm. The two firms engage in Bertrand competition, and as a result, the incumbent firm retains the worker offering the new contract $V_{e,t}^*(y')$.

3. $V_{e,t}^*(y) \leq V_{e,t}^*(y')$; in this case the poaching firm has a willingness to pay that is no less than the incumbent's. If this condition holds with strict inequality, the current match is terminated and the worker is poached at the maximum value of the contract that the incumbent is willing to pay. If instead the worker is poached by a firm with equal productivity, it is assumed that job switching takes place with probability ν. In either case, the continuation value of the contract obtained by the worker is $V_{e,t}^*(y)$.

The bargaining protocol above, together with the assumption that entrant firms make zero profits in expectations, yields the free entry-condition, i.e., equation (12) in the main text, which we display again below for convenience:

$$\begin{aligned} c^f + \frac{c}{\varpi_t} &= \frac{u_{0,t}}{u_{0,t} + s_t(1-u_{0,t})} \left\{ \xi_b S_t(y_b) + \xi_g S_t(y_g) \right\} \\ &\quad + \frac{s_t(1-u_{0,t})}{u_{0,t} + s_t(1-u_{0,t})} \left\{ \xi_g \frac{l_{b,t}^0}{1-u_{0,t}} \left[S_t(y_g) - S_t(y_b) \right] \right\}. \end{aligned} \tag{38}$$

Substituting out for the surplus functions in the above equations requires some steps. Start by considering the case of a firm that has promised to pay the contract $\{w_s^*(y)\}_{s=t}^{\infty}$, which implies that the firm breaks even in expectation and is not able to promise higher wage payments in case it enters an auction with a poaching firm. In this case, if no outside offers arrive the worker receives a continuation value of $V_{e,t}^*(y)$ from the incumbent firm. Otherwise the worker is poached and, in accordance with point (3) above, receives from the new firm a contract that is also worth $V_{e,t}(w', y') = V_{e,t}^*(y)$. So either way, the worker receives a contract of value $V_{e,t}^*(y)$. The value to a worker of being employed under the contract $\{w_s^*(y)\}_{s=t}^{\infty}$ can therefore be written as:

$$V_{e,t}^*(y) = \varphi_t y + \beta E_t \frac{\lambda_{t+1}}{\lambda_t} \left[\delta V_{u,t} + (1-\delta) V_{e,t+1}^*(y)\right], \tag{39}$$

where $\varphi_t y$ is the marginal revenue product of selling y units of the service to the price setters. Subtracting (37) from the above equation yields:

$$V_{e,t}^*(y) - V_{u,t} = \varphi_t y - \frac{b}{\lambda_t} + (1-\delta)\beta E_t \frac{\lambda_{t+1}}{\lambda_t} \left[V_{e,t+1}^*(y) - V_{u,t+1}\right]. \tag{40}$$

Notice that the value to the worker of extracting all the rents associated with a type-y match, $V_{e,t}^*(y) - V_{u,t}$, is in fact simply the surplus $S_t(y)$. Iterating forward on the above expression, we can define the surplus of a match with productivity y as:

$$S_t(y) = E_t \left[\sum_{\tau=0}^{\infty} (1-\delta)^{\tau} \left(\frac{\lambda_{t+\tau}}{\lambda_t} \varphi_{t+\tau} y - \frac{b}{\lambda_t}\right)\right]. \tag{41}$$

Notice that the surplus function above is affine increasing in y, which implies that firms with higher productivity win the auction and, therefore, workers cannot move to jobs with lower productivity. For convenience, we can rearrange the above expression as

$$S_t(y) = y\mathcal{W}_t - \frac{b\lambda_t^{-1}}{1-\beta(1-\delta)}, \tag{42}$$

where

$$\mathcal{W}_t = \varphi_t + \beta(1-\delta) E_t \frac{\lambda_{t+1}}{\lambda_t} \mathcal{W}_{t+1}. \tag{43}$$

Seen from the point of view of a service sector firm, $\mathcal{W}_t$ can be interpreted as the expected present discounted value of the entire stream of current and future real marginal revenues derived from selling one unit of the service until separation. From the point of view of a price setting firm, which purchases labor services, $\mathcal{W}_t$ can be interpreted as the expected present discounted value of the cost of purchasing one unit of the labor service by a firm until separation.

Using equation (42) we can now substitute for the surplus functions and rearrange to rewrite

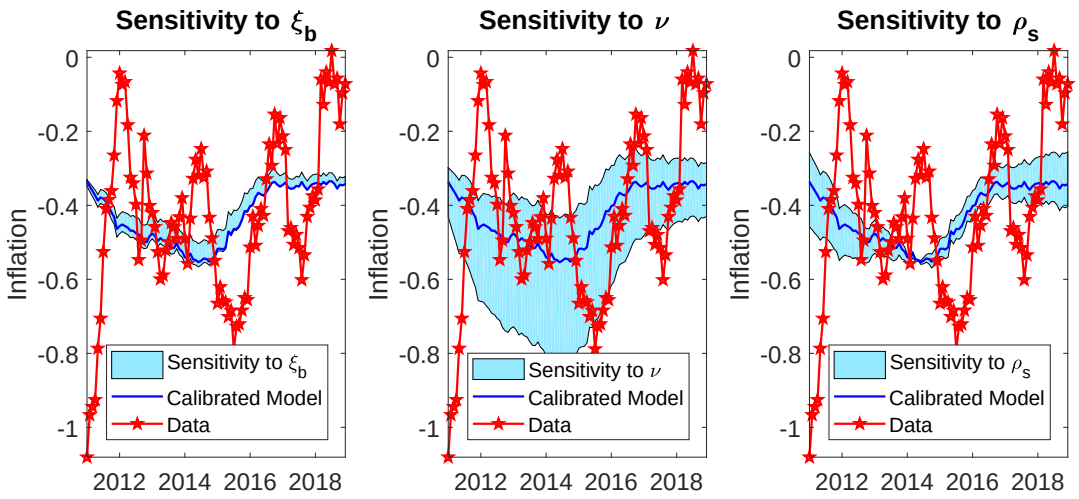

Figure 12: Robustness. Left panel: The shaded area show the sensitivity of the model's predicted year-over-year inflation rate to changes in the probability that the meeting between the worker and the firm generates a bad match (ξ_b). The blue solid line denotes the model's predicted year-over-year inflation rate for our baseline calibration shown in Table 1. The red starred line denotes the year-over-year inflation rate in the data (core inflation according to the Price Index for Personal Consumption Expenditures, or PCE) in deviations from the Survey of Professional Forecasters' PCE inflation expectations over the next ten years. The middle and right panels show the same plot when we perturb the probability that workers accept an offer if they are indifferent (ν) and the persistence of the on-the-job search rate (ρ_s).

the free-entry condition (12) as:

$$c^f + \frac{c}{\varpi_t} = \frac{u_{0,t}}{u_{0,t} + s_t\left(1 - u_{0,t}\right)}\left[\mathcal{W}_t\left(\xi_b y_b + \xi_g y_g\right) - \frac{b\lambda_t^{-1}}{1 - \beta\left(1 - \delta\right)}\right] + \frac{s_t}{u_{0,t} + s_t\left(1 - u_{0,t}\right)}\xi_g l_{b,t}^0 \mathcal{W}_t\left(y_g - y_b\right). \tag{44}$$

G Robustness

The shaded area in the panels of Figure 12 shows how the model's prediction of inflation changes as we vary the probability of meeting a worker that is a bad match for the firm ξ_b (left), the probability that workers switch jobs if they receive an outside offer that makes them indifferent (ν) (middle), or the persistence of the on-the-job search rate (ρ_s) (right). We consider values of the parameter ξ_b ranging from 0.6 through 0.8, values of the parameter ν ranging from 0.25 through 0.75, and values of the parameter ρ_s ranging from 0 through 0.97 (the highest confidence bound when the AR parameter of the series of the-on-the-job search $\tilde{s}_t$ is estimated by OLS). The blue solid line and the red starred lines denote the model's predicted inflation rate and the core PCE inflation gap for the baseline calibration reported in Table 1, respectively. These lines are the same as the ones plotted in the left panel of Figure 5.

H Solving the Model with the ZLB Constraint

After being solved, our linearized model with the occasionally binding ZLB constraint in equation (20) can be represented in state-space form as follows:

$$s_t = \Gamma_0 s_{t-1} + \Gamma_1 \varepsilon_t^1 + \Gamma_2 \varepsilon_t^2 \tag{45}$$

where the first $k+1$ rows of s_t contain the current policy rate and the expectations of the policy rate in quarter $t+1, ..., t+k$. The model's structural shocks are contained in ε_t^2. This vector of shocks includes the preference shock and the shocks to the on-the-job search rate. The linear system above also features a vector of dummy shocks ε_t^1. These shocks in ε_t^1 are appended to the Taylor rule so that the constrained Taylor rule in equation (20) can be written as

$$\frac{R_t}{R^*} = \left(\frac{R_{t-1}}{R^*}\right)^{\rho_r} \left[\left(\frac{\Pi_t}{\Pi^*}\right)^{\phi_\pi} \left(\frac{Q_t}{Q^*}\right)^{\phi_y}\right]^{1-\rho_r} + \sum_{j=0}^{k} \eta_{t-j}^j, \tag{46}$$

where η_t^j are $k+1$ monetary shocks that are known by agents at time t and will hit the economy at time $t+j$. These shocks belong to the vector ε_t^1 in equation (45). These dummy shocks serve the sole purpose of enforcing the ZLB constraint (i.e., prevent agents from expecting negative nominal interest rates in any state of the world). Thus, the realizations of these dummy shocks will be equal to zero in every states of the world in which the current and expected nominal interest rates do not violate the ZLB constraint. It should be noted that the matrix Γ_1 is a matrix with $k+1$ columns.

As explained in the main text, the shocks are obtained by inverting the 2×2 square matrix $Z\Gamma_2$, where the matrix Z is a 2×2 observation matrix such that $Y_t = Z s_t$ with the vector Y_t including the observables (i.e., the unemployment rate and the EE flow rate) used in the empirical exercise whose results are described in Section 4.3 and Section 4.5. Under the assumption that the matrix $Z\Gamma_2$ is invertible (as it is in our case), this inversion allows us to retrieve the sequence of shocks ε_t^2 that identically explains the observed rate of unemployment and the EE rate.

We start by setting $t = 1$, which denotes the first period of our sample Y_t, and go through the following steps:

1. Given the realization of the two shocks ε_t^2 at time t, we set the matrix $\Psi(0) = \mathbf{0}_{k+1 \times k+1}$, $\varepsilon_t^1(0) = \mathbf{0}_{k+1 \times 1}$, $i = 0$, and go to Step 2.

2. Define the vector of adjustments to forward guidance shocks $\Delta\varepsilon_t^1$ that ensures the current

and/or the expected path of the future interest rates will respect the ZLB as follows:

$$\Delta\varepsilon_t^1 = \left(\Gamma_1^{(0:k)}\right)^{-1}\left[-\ln R_* - \Gamma_0^{(0:k)} s_{t-1} - \Gamma_1^{(0:k)}(i)\cdot\Psi(i)\,\varepsilon_t^1(i) - \Gamma_2^{(0:k)}\varepsilon_t^2\right], \tag{47}$$

where $\Gamma_1^{(0:k)}$ denotes the square submatrix made of the first $k+1$ rows of the matrix Γ_1. With $\Delta\varepsilon_t^1$ at hand, we update $\varepsilon_t^1(i+1) = \varepsilon_t^1(i) + \Delta\varepsilon_t^1$. Note that if the ZLB constraint is not binding at time t, $\Delta\varepsilon_t^1 = \mathbf{0}_{k+1\times 1}$.

3. Check if the below inequality is satisfied (the ZLB is not binding),

$$\Gamma_0^{(0:k)} s_{t-1} + \Gamma_1^{(0:k)}\cdot\Psi(i)\,\varepsilon_t^1(i+1) + \Gamma_2^{(0:k)}\varepsilon_t^2 > -\ln R_*. \tag{48}$$

We adjust the diagonal matrix of zeros and ones, $\Psi(i+1)$, so that the set of horizons at which the ZLB is binding is characterized with a value equal to one in this matrix. If $\Psi(i+1)\,\varepsilon_t^1(i+1) \neq \Psi(i)\,\varepsilon_t^1(i)$, set $i = i+1$ and go to Step 2, or else the fixed point is found and we set $\varepsilon_t^1 = \Psi(i+1)\,\varepsilon_t^1(i+1)$ and go to Step 4.

4. Compute the next period's state vector as follows:

$$s_t = \Gamma_0 s_{t-1} + \Gamma_1\varepsilon_t^1(i+1) + \Gamma_2\varepsilon_t^2. \tag{49}$$

Set $t = t+1$, and go back to Step 1.

The s_t coming from equation (49) is the vector containing the model-predicted values of the state variables, which is used to generate all the empirical results of the paper.

I Propagation of Preference Shocks

Figure 13 shows the responses to a negative preference shock. As done in the main text, we report the responses of the labor market variables (unemployment, bad matches, and good matches) at the beginning of the period, and as such they do not respond on impact by construction. When the preference shock hits, households want to save more and consume less. As a result, households' demand for the differentiated goods falls, leading to a drop in the price setters' demand for the labor service and hence in its relative price φ_t. Forward-looking price setters anticipate that marginal costs will remain low and cut their price, leading the inflation rate to fall. Concurrently, the weakening of the price setters' demand for labor service reduces entry in the labor market, which in turn induces unemployment to rise over the subsequent periods. As the fraction of unemployed job seekers surges, labor becomes cheaper in expectation for an entrant service firm, because it is now more likely to extract a nonzero surplus from

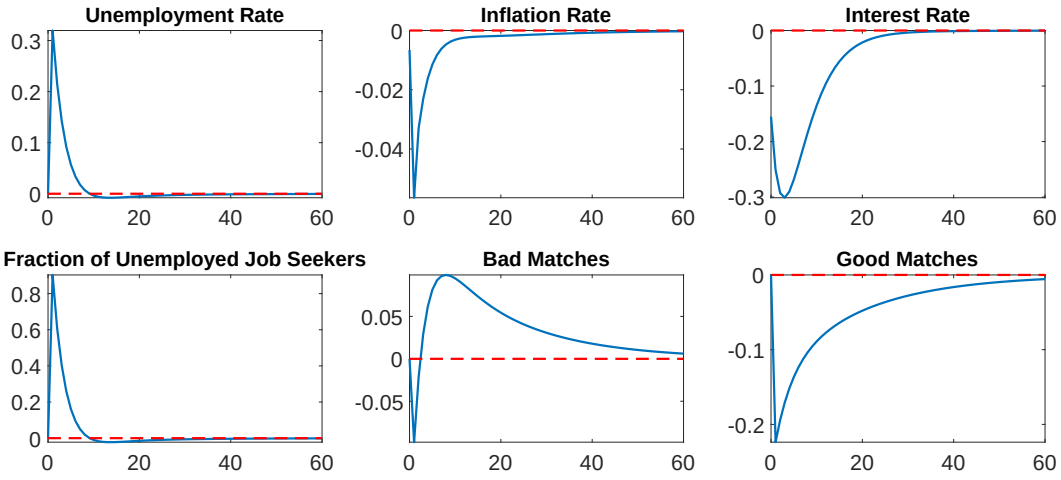

Figure 13: Impulse responses to a negative preference shock by one standard deviation. The unemployment rate, the fraction of unemployed job seekers, and the shares of good and bad matches are measured at the beginning of the period, consistently with the definition of labor market slack Σ_t in equation (15). Units: percentage points. Inflation and interest rate are expressed in annualized rates.

the match. As a result, equation (15) implies an increase in labor market slack, along with a decrease in the price of labor service, and, therefore, an even further drop in inflation in the second period.

Also note that the stock of bad matches falls initially and then rises as the entry of more labor service firms allows unemployed workers to find jobs and thus climb the ladder anew. This rise in bad matches, along with the fall in good matches, further contributes to keeping labor cheap for longer and to depressing price dynamics. Our measure of slack captures these effects through the second term on the right-hand side of equation (15).

J The Model with Endogenous Search Intensity

Assume that each period, every employed worker draws a fixed cost of search from a uniform distribution

$$g(\varsigma) \sim \mathcal{U}[\xi_t \varsigma, \xi_t \varsigma + \varsigma], \tag{50}$$

where $\varsigma > 0$ is a parameter. We assume that the aggregate shock to the cost of searching on the job behaves as follows

$$\xi_t = (1 - \rho_\xi)\bar{\xi} + \rho_\xi \xi_{t-1} + \varepsilon_{\xi,t}, \quad \varepsilon_{\xi,t} \backsim \boldsymbol{N}(0, \sigma_\xi). \tag{51}$$

Let $l_t^{i,u}$ denote the number of workers employed in matches of type $i \in (b, g)$ under zero surplus, and $\bar{\varsigma}_t^{i,u}$ the threshold value that makes them indifferent to search on the job or not. Similarly, let $l_t^{i,j}$ and $\bar{\varsigma}_t^{i,j}$ denote the measure of workers and threshold search costs that refer

to workers employed in matches of type $i \in (b,g)$ under extraction of surplus of a job of type $j \in (b,g)$. The threshold value of the search cost that makes a worker employed in a bad match under zero surplus indifferent between searching on the job and not searching is $\bar{\varsigma}_t^{b,u} = \phi_t S_{b,t}$. For a worker employed in a good match under zero surplus, the threshold is $\bar{\varsigma}_t^{g,u} = \phi_t \left(\xi_b S_{b,t} + \xi_g S_{g,t}\right)$. For a worker employed in jobs of type i under full extraction of a bad job surplus we get the following threshold: $\bar{\varsigma}_t^{b,b} = 0$. For worker employed in a good match under partial extraction, we obtain $\bar{\varsigma}_t^{g,b} = \phi_t \xi_g \left(S_{g,t} - S_{b,t}\right)$. Finally, for workers employed in good jobs under full extraction of the surplus, the threshold for searching is $\bar{\varsigma}_t^{g,g} = 0$.

The assumption that $g(\varsigma)$ is uniformly distributed implies that

$$Prob\left\{\varsigma_{j,t} < \bar{\varsigma}_t^{i,j}\right\} \equiv G\left(\varsigma < \bar{\varsigma}^{i,j}\right) = \frac{\bar{\varsigma}_t^{i,j}}{\varsigma} - \xi_t, \tag{52}$$

if $\xi_t \varsigma < \bar{\varsigma}_{j,t} < \xi_t \varsigma + \varsigma$. Note that $G\left(\varsigma < \bar{\varsigma}^{i,j}\right)$ for each type (i,j) represents the fraction of workers whose period cost of search is lower than their threshold. By the law of the large numbers, this value equals the measure of workers of type (i,j) who search on the job. These measures are necessary to characterize the workers' laws of motion across the rungs of the ladder.

The laws of motion for the workers employed under zero surplus is

$$l_t^{i,u} = l_{0,t}^{i,u} + \phi_t \xi_i u_{0,t} - l_{0,t}^{i,u} G\left(\bar{\varsigma}_t^{i,u}\right) \phi_t \quad \text{for } i = \{b,g\}, \tag{53}$$

$$l_{0,t+1}^{i,u} = (1-\delta)\, l_t^{i,u}. \tag{54}$$

The law of motion for the workers employed in bad matches under full extraction of bad match surplus is

$$l_t^{b,b} = l_{0,t}^{b,b} + l_{0,t}^{b,u} G\left(\bar{\varsigma}_t^{b,u}\right) \xi_b \phi_t - l_{0,t}^{b,b} G\left(\bar{\varsigma}_t^{b,b}\right) \phi_t \xi_g \tag{55}$$

$$l_{0,t+1}^{b,b} = (1-\delta)\, l_t^{b,b}.$$

The law of motion for the workers employed in good matches under full extraction of bad match surplus is

$$l_t^{g,b} = l_{0,t}^{g,b} + l_{0,t}^{b,u} G\left(\bar{\varsigma}_t^{b,u}\right) \xi_g \phi_t + l_{0,t}^{b,b} G\left(\bar{\varsigma}_t^{b,b}\right) \phi_t \xi_g + l_{0,t}^{g,u} G\left(\bar{\varsigma}_t^{g,u}\right) \xi_b \phi_t - l_{0,t}^{g,b} G\left(\bar{\varsigma}_t^{g,b}\right) \phi_t \xi_g \tag{56}$$

$$l_{0,t+1}^{g,b} = (1-\delta)\, l_t^{g,b}. \tag{57}$$

Equations (53), (55) and (56) above solve for $l_t^{i,u}$ and $l_t^{i,b}$, for $i = \{b,g\}$.

The total measure of workers searching at every point in time is:

$$l_{0,t}^{s} = l_{0,t}^{b,u} G\left(\bar{\varsigma}_t^{b,u}\right) + l_{0,t}^{g,u} G\left(\bar{\varsigma}_t^{g,u}\right) + l_{0,t}^{b,b} G\left(\bar{\varsigma}_t^{b,b}\right) + l_{0,t}^{g,b} G\left(\bar{\varsigma}_t^{g,b}\right). \tag{58}$$

The law of motion for the workers employed in good matches under full extraction of the good job surplus is:

$$l_t^{g,g} = l_{0,t}^{g,g} + \left[l_{0,t}^{g,u} G\left(\bar{\varsigma}_t^{g,u}\right) + l_{0,t}^{g,b} G\left(\bar{\varsigma}_t^{g,b}\right) \right] \xi_g \phi_t, \tag{59}$$

$$l_{0,t+1}^{g,g} = (1-\delta)\, l_t^{g,g}. \tag{60}$$

Total employment can be computed as

$$l_{0,t} = l_{0,t}^{b,u} + l_{0,t}^{b,b} + l_{0,t}^{g,u} + l_{0,t}^{g,b} + l_{0,t}^{g,g}. \tag{61}$$

K The Evolution of Match-Quality in the Model with Endogenous Search

The shock ξ_t is pinned down by the joint dynamics of the EE rate and the UE rate implied by the observed unemployment rate. The lower degree of procyclicality of the EE rate relative to the job finding rate leads the shock ξ_t to lower the support of the distribution of search costs in recession and to raise it in expansion. A lower support implies that, everything else equal, more employed workers will draw search costs that lie below the threshold that makes them indifferent between searching or not. Hence, more of them will search. How the probability to search on the job responds to the shock ξ_t depends on the position of workers on the ladder and the surplus they are able to extract. Specifically, the fraction workers employed in a bad match is very countercyclical and attains its historical low at the end of the sample.[35] This implies that the bad matches remain persistently above their long-run level throughout the last recovery. Remarkably, even though the job finding rate attains very high levels at the end of the sample, our model with endogenous on-the-job search predicts that the mass of bad matches decreases only very slowly. This sluggish adjustment in the measure of bad matches is mainly accounted for by the workers who are stuck in a bad match under full surplus extraction. A graph showing the breakdown in the behavior of bad matches by surplus extraction is provided in Figure 14.

The left plot of Figure 14 shows the decomposition of bad matches into workers in bad matches with no surplus (the blue solid line) and those with full surplus extraction of surplus (the black dashed line). The latter are workers in a bad match that were able to secure an outside offer from another firm that is a bad match for the worker. As one can see, the share

[35] Workers who have not received any outside offers and thereby receive no surplus are the ones who search more intensively. However, their propensity to search is highly procyclical being primarily affected by the dynamics of the job finding rate. At the end of the sample, these workers search a lot A similarly procyclical pattern is followed by the search rate of those workers employed in good matches under partial extraction of surplus (i.e., they are in a good match but they have so far failed to secure a good outside offer). Nevertheless, their search rate is less volatile than those of the workers who are extracting zero surplus.

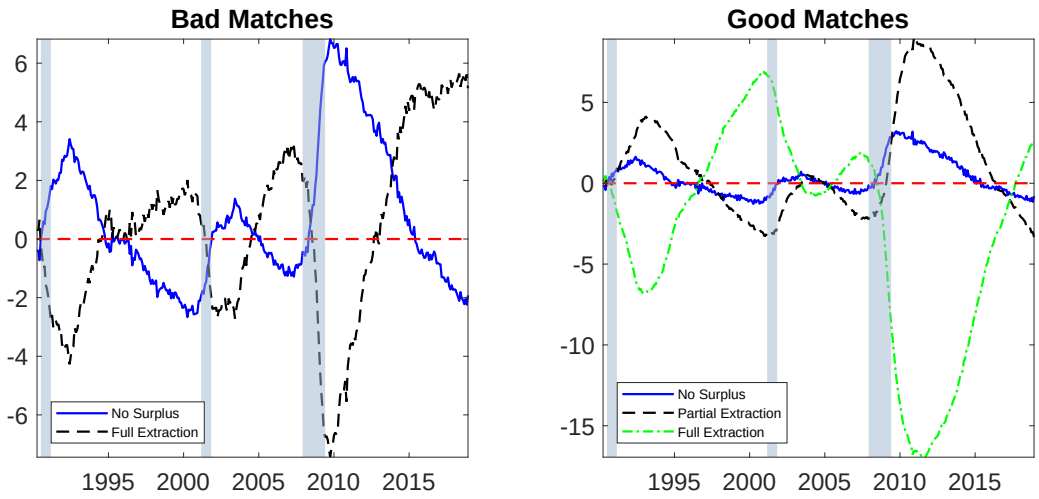

Figure 14: Breakdown of Bad Matches and Good Matches. Left plot: bad matches. Right plot: good matches. Unit: deviations from steady state in percentage points.

of these workers raises at the beginning of the expansion, which is typical of the job ladder, but, instead of converging back to steady state, it levels out until the end of the expansion. This happens because the propensity of these workers to search on the job declines during the recovery.

The presence of workers who are stuck in a bad match, having failed to attract a good offer, contributes to mitigate the intensity of interfirm wage competition. As explained in Section 3.6 in the context of the baseline model, these workers represent a relatively cheap source of labor from the perspective of an entrant firm, given their expected productivity. As a result, this large share of bad matches brings about slack, thereby exerting downward pressure on inflation. Indeed, the model with endogenous on-the-job search predicts inflation to remain below its longer-run level (2%) throughout the long post-Great Recession recovery, echoing the results obtained in the baseline model in Section 4.