Chapter 17

LABOR-MARKET FRICTIONS AND EMPLOYMENT FLUCTUATIONS

ROBERT E. HALL

Hoover Institution and Department of Economics, Stanford University

National Bureau of Economic Research

Contents

Abstract 1138
Keywords 1138
1. Introduction 1139
2. The baseline neoclassical model 1140
  2.1. Failure of amplification in the baseline neoclassical model 1141
  2.2. Evidence about technology impulses 1141
  2.3. Failure of the baseline neoclassical model to explain persistence 1142
  2.4. Absence of unemployment from the baseline neoclassical model 1143
3. Amplification 1145
  3.1. Elastic conventional labor supply 1145
  3.2. Empirical research 1148
  3.3. Unemployment 1150
    3.3.1. Mechanism design and labor contracts 1154
    3.3.2. The modern strategic view of the employment relationship 1157
    3.3.3. Efficiency wages 1157
    3.3.4. Job destruction 1158
    3.3.5. Reorganization and reallocation 1160
4. Persistence 1162
  4.1. Time-consuming matching in the labor market 1162
  4.2. The importance of secondary job loss for persistence 1163
5. Conclusion 1167
Acknowledgments 1167
References 1167

Handbook of Macroeconomics, Volume 1, Edited by J.B. Taylor and M. Woodford
© 1999 Elsevier Science B.V. All rights reserved

Abstract

The labor market occupies center stage in modern theories of fluctuations. The most important phenomenon to explain and understand in a recession is the sharp decline in employment and jump in unemployment. This chapter considers explanations based on frictions in the labor market. Earlier research within the real business cycle paradigm considered frictionless labor markets where fluctuations in the volume of work effort represented substitution by households between work in the market and activities at home. A preliminary section of the chapter discusses why frictionless models are incomplete – they fail to account for either the magnitude or persistence of fluctuations in employment. And the frictionless models fail completely to describe unemployment. The evidence suggests strongly that consideration of unemployment as a third use of time is critical for a realistic model. The two elements of a theory of unemployment are a mechanism for workers to lose or leave their jobs and an explanation for the time required from them to find new jobs. Theories of mechanism design or of continuous re-bargaining of employment terms provide the first. The theory of job search together with efficiency wages and related issues provides the second. Modern macro models incorporating these features come much closer than their predecessors to realistic and rigorous explanations of the magnitude and persistence of fluctuations.

Keywords

JEL classification: E24

1. Introduction

The bulk of modern fluctuations theory fits into the broad framework of impulses–amplification–persistence. In this framework, fluctuations begin with a random impulse. These can be shifts in technology or preferences, shifts in monetary or fiscal policy, or spontaneous movements in consumption, investment, or other components of spending. Most impulses of realistic size have small effects on employment in a standard neoclassical model. To explain the observed volatility of employment and other important aggregates, some form of amplification must occur. Modern thinking about sources of amplification has focused primarily on highly elastic labor supply. Labor supply is elastic because workers have economically valuable alternative uses of their time. Though non-market alternatives lie behind the standard view of labor supply and the view of labor supply in the original real business cycle model, the alternative that is most prominent in recent models is job search. In fact, more than anything else, this chapter is about the successful integration of unemployment theory into formal dynamic general-equilibrium models. Not only do the new general-equilibrium models comprehend the important phenomenon of unemployment, but it turns out that unemployment is key to amplification and persistence.

Without consideration of unemployment, earlier dynamic general-equilibrium models explained persistence in employment largely through persistence in driving forces. The models themselves did not contain much in the way of persistence mechanisms. Where unemployment is considered explicitly, persistence arises naturally from the time-consuming process of placing unemployed workers in jobs following an adverse impulse.

This chapter does not consider the origins of impulses, nor the propagation of fluctuations across industries, though that propagation may interact with amplification and persistence. The chapter focuses on frictions in the labor market, some of which amplify impulses and others of which result in persistence, especially of unemployment.

Most of the research I will consider here is developed within dynamic stochastic general equilibrium (DSGE) models. By placing amplification and persistence mechanisms in formal general equilibrium models, contributors to modern fluctuations research achieve a degree of clarity missing from earlier macroeconomics. Confusing notions from earlier work, such as "aggregate demand" and "supply shocks," are giving way to clearer general equilibrium counterparts, such as monetary and fiscal impulses and shifts in the terms of trade. Moreover, research in the impulses-amplification-persistence mode has come close to eliminating the traditional polarization of macro researchers. Impulses come from technology, policy, and spontaneous shifts, a list broad enough to include almost any earlier idea about the sources of fluctuations. Old-fashioned Keynesian ideas such as wage rigidity, new Keynesian ideas such as efficiency wages, and ideas about imperfect information all compete on an equal footing to explain amplification, and are no longer assigned to warring schools of thought.

2. The baseline neoclassical model

For several decades, a baseline neoclassical model has anchored macroeconomics. Because the baseline model is successful neither as a model of fluctuations nor of growth, both of the major branches of the field have explored alternatives to the baseline. For the purposes of this chapter, it is useful to lay out the baseline model and explain its failure as a model of fluctuations. Lack of amplification of impulses and lack of persistence of the resulting responses are the symptoms of the failure. In the baseline model, workers choose between two activities, work and leisure; there is no consideration of unemployment.

The baseline model is so well known that it will suffice to describe it mainly verbally – see Campbell (1994) and Romer (1996, chapter 4) for more extensive discussions along this line. A single type of output is consumed or invested. It is produced by a Cobb–Douglas technology with constant returns to scale. Labor and capital are the inputs. Consumers' preferences are ordered by the time integral of log consumption plus weighted log leisure – that is, the intertemporal utility function is

$$\int_0^\infty e^{-\beta t} [\log c(t) + \lambda \log(\bar{n} - n(t))] \, dt. \tag{2.1}$$

Log consumption insures that static labor supply has zero wage elasticity – models typically adopt this specification to match the positive trend in real wages to the zero trend in annual hours per worker[1].

In principle, the baseline model could be driven by almost any kind of impulse – shifts in technology or preferences, changes in policy regimes, or random elements of policy. Kydland and Prescott's (1982) pioneering exploration of fluctuations using the baseline model and alternatives focused on vibrations of the aggregate production function as the single driving force. Much of the ensuing literature retained that focus, though changes in government purchases – an important topic before Kydland and Prescott's formalization of the baseline model – remain a second important driving force in that literature [see, for example, Aiyagari, Christiano and Eichenbaum (1992)].

Empirical measures of aggregate technology – generally obtained by calculating the Solow (1957) residual – suggest that changes in technology are quite persistent. In fact, a random walk is not a bad approximation to the stochastic properties of aggregate technology or for the technologies of particular industries. To put it differently, the year-to-year Solow residual, which measures the change in technology, is close to white noise. The custom has developed in the literature on DSGE models to model the stochastic process of aggregate technology as first-order autoregressive with a serial

[1] See King and Rebelo (1999) for further discussion of this point.

correlation parameter of 0.95. In that case, an innovation in technology in one period is followed by small movements in the opposite direction for many succeeding periods, as the level of technology gradually (at a rate of 5 percent per period) returns to normal.

In this setting, a rough description of the perturbation in general equilibrium following, say, a negative shock to technology is the following: Immediately, employment falls. Output falls both because of the direct effect of the decline in productivity and because of the decline in employment. Since the capital stock is unaffected, the marginal product of capital and thus the interest rate fall. Consumption rises and investment declines by more than the decline in output. There follows an extended period of gradual rises in the interest rate and real wage back to their normal levels. As a result, consumption – governed by its Euler equation – falls back to normal and, similarly, employment rises back to normal.

2.1. *Failure of amplification in the baseline neoclassical model*

As I noted earlier, a central problem of fluctuations theory is that small impulses seem to result in large movements, especially those we call recessions. Can the baseline neoclassical model explain the actual magnitude of observed fluctuations based on the likely magnitude of technology shifts? From the starting point in Kydland and Prescott (1982), the answer has been no. Some form of amplification beyond what is found in the baseline model is needed.

Campbell (1994) investigates this issue carefully. His Table 3 shows that a 1-percent decline in technology that is reversed at 5 percent per quarter lowers employment by only 0.45 percent in the baseline model. In other words, to explain the 3-percent decline in employment typical of a recession, the model has to invoke a decline in technology of about 7 percent! The challenge to fluctuations theory is to change the baseline model in a reasonable way that overcomes this small response. The next section considers a number of attempts along this line.

2.2. *Evidence about technology impulses*

Empirical work shows that the standard measure of technology shifts – the Solow residual -- is correlated somewhat positively with employment and more positively with output. For example, in Hall (1997), the correlation of the Solow residual with hours of work at business-cycle frequencies is 0.32 while the correlation with real GDP is 0.75. The interpretation of the correlation is a disputed issue, however. A path of research whose origin was wholly apart from measuring macro impulses has suggested that the correlation is the result of the failure of assumptions underlying Solow's method. Hall (1988, 1990) observed that one form of the Solow residual could be used to test the assumption of competition in output markets and to measure the ratio of price to marginal cost. The resulting corrected Solow residual is hardly correlated with employment. Another form of the Solow residual can be used to test the assumption of constant returns to scale and to measure an index of increasing returns. Again, the

Solow residual corrected for increasing returns is less correlated with output and is hardly correlated with employment.

A substantial subsequent literature has questioned Hall's finding of high markups matched by equally high returns to scale. New work has substituted better measures of labor and capital input [Basu (1996), Burnside, Eichenbaum and Rebelo (1993)] and corrected aggregation bias [Basu and Fernald (1997)]. With all corrections in place, there is no remaining evidence of a correlation of the Solow residual with employment. Thus, there are two reasons to question the existence of technology shifts that are correlated with employment changes. First, what appears to be shifts in the simple Solow residual are actually artifacts of imperfect competition or increasing returns. Second, the correlation is the result of errors of measurement. Either finding is troublesome for the view that technology shocks are an important driving force.

Hall (1997) investigates the role of the technology impulse in a semi-econometric general equilibrium framework. The model adopts the premise of the baseline model mentioned in the previous section – by making the kernel of utility depend on the log of consumption, the model excludes any direct effect of a technology shift on labor supply. A direct effect is one that occurs even in an economy without capital or other methods for shifting resources between time periods. The absence of direct effects occurs when the income and substitution effects in the corresponding static labor supply function offset each other.

Hall observes that all of the effects of the technology impulse on employment must operate through the intertemporal channel, that is, through investment – a positive technology innovation sets off an investment boom. Hours of work rise, GDP rises, and consumption falls, as the economy moves quickly to take advantage of higher productivity. Hall examines the empirical relation between the investment/GDP ratio and the simple uncorrected Solow residual. He finds a robust positive relation – a one-percent shift of the production function causes about a one-percent increase in the investment/GDP ratio. But the fraction of employment volatility explained by the technology impulse is essentially zero. Further, the use of a more refined version of the Solow residual would probably eliminate what little role the technology impulse is found to have.

2.3. *Failure of the baseline neoclassical model to explain persistence*

Cogley and Nason (1995) observed that the pattern of movements of employment and output in the baseline neoclassical model is essentially that of its technology driving force. The point would apply to other driving forces as well. The neoclassical model cannot mimic the pattern of recession and recovery in response to a single shock, despite the common-sense impression that recessions often result from discrete shocks. In particular, after a one-year temporary decline in productivity, employment and output return to normal immediately in the neoclassical baseline model.

2.4. *Absence of unemployment from the baseline neoclassical model*

The baseline model considers only two uses of time – employment and leisure. A strong consensus has emerged in macroeconomic thinking that a realistic model needs to consider a third use of time – job search or unemployment. Hall (1997) takes the following approach to demonstrating the need for explicit consideration of unemployment: He considers a neoclassical model without unemployment, but one where shifts in household preferences drive fluctuations along with shifts in technology and changes in government purchases. Because he considers three aggregate variables – output, consumption, and hours of work – he is able to solve for the values of the three impulses from the values of the three observed variables, based on standard values for the parameters of the neoclassical model. Almost all the explanatory power is assigned to the preference shift. Changes in government purchases have a small role because the observed changes are small and not generally associated with booms or recessions. Technology shifts also receive little weight because they operate solely through the intertemporal investment channel and because they should cause employment to change much more than consumption.

Hall models the preference shift as a random variable that determines the marginal rate of substitution contemporaneously between consumption and leisure. Shifts in the variable cause employment and consumption to move in the same direction. Hall uses an empirical approach to determine the relative explanatory powers of the driving forces for employment fluctuations. Almost all the credit goes to the preference shift. But his conclusion is not that preference shifts are actually a major driving force. Rather, another use of time – unemployment – is left out of the model. Periods of higher unemployment are times when employment and consumption are both low. A better way to explain the positive correlation of employment and consumption at business-cycle frequencies is to bring unemployment explicitly into the model.

Rotemberg and Woodford (1996) demonstrate the failure of the baseline model in a rather different way. They focus on the joint time-series properties of employment, output, and consumption. They show that the data contain a business cycle in the weak sense defined by Beveridge and Nelson (1981) – the data tend to return to a long-run trend whenever they deviate from that trend in the short run. The forecast of employment growth is unusually high, for example, if employment is below its trend because a recession occurred recently. In the baseline model driven by a technology shock that follows a strict random walk [not the AR(1) process with a serial correlation of 0.95 I discussed earlier], employment and other variables lack almost any tendency to return to normal. They derive measures of that tendency from a 3-variable vector autoregression. When there is a tendency to return to normal, the VAR forecasts future values for the variables that are different from the current values. They measure this forecasting power at various horizons. Their findings are summarized in Table 1.

In the baseline model, the current value of output is close to the best value of the forecast of future values at any horizon. There is a very slow-moving forecastable component in the baseline model, associated with capital accumulation, so the standard

Table 1
Failure of the baseline model[a]

	Standard deviation of forecastable component of output (percent)			
	8 quarters ahead	12 quarters ahead	24 quarters ahead	Infinitely ahead
Baseline model	0.17	0.23	0.36	0.53
Actual data with VAR	2.95	3.22	3.05	3.06

[a] Data from Rotemberg and Woodford (1996).

deviation of the forecastable component rises as the horizon lengthens. In the actual data for the USA, there is a pronounced rebound from abnormal conditions in a year or two. The standard deviation of the change in output forecasted by the VAR is about 3 percent at all horizons. In addition to being much larger than the forecastable component in the baseline model, the time profile of the forecasting power is quite different in the actual US economy – the forecasted change occurs almost entirely in the first 8 quarters. In the baseline model, the forecasting power grows quite a bit after 8 quarters.

Related failures of the baseline model are revealed in the correlations of the forecastable components. In the baseline model, the forecasted change in work effort should have the opposite sign from the forecasted change in output. When a shock has caused the economy to be at a point below its steady-state capital stock, work effort will be above its steady state while output will be below. As capital is accumulated, output rises and work effort falls. In fact, forecasted movements in output and work effort are in the *same* direction. After a recession, both hours of work and output rise more rapidly than normal.

It appears that a reasonable explanation for the failure of the baseline model in Rotemberg and Woodford's work is the absence of unemployment in the model. The forecastable rebound that occurs in the US economy following a recession occurs during the period when workers displaced during the recession are making their way back into long-term employment. In the baseline model, there is no burst of unemployment in the first place and no two-year period of rematching. Both anomalies reported by Rotemberg and Woodford are resolved by adding unemployment to the model – the augmented model has much more predictable recovery from bad (or good) shocks, and output and work effort move in the same direction during recoveries.

Not only does consideration of unemployment provide a more sensible interpretation of correlations among key macro variables, but modern ideas about both amplification and persistence often involve job destruction and job search, key ideas in the modern theory of unemployment.

3. Amplification

Amplification occurs when the response of employment to a driving force is stronger than in the baseline neoclassical model. Macro research in the DSGE framework has recognized the need for amplification mechanisms since Kydland and Prescott's (1982) paper launched the framework. The mechanisms I discuss in this section all involve elastic labor supply, either in the conventional sense or in the sense that there is another activity – job search – that is a substitute for work.

3.1. Elastic conventional labor supply

The earliest amplification mechanisms invoked elastic labor supply in the standard setting where workers choose between work effort and leisure. If labor supply is more elastic – for example, if the labor part of the kernel of the utility function is $(\bar{n} - n)^{0.8}$ instead of $\log(\bar{n} - n)$ – the response of employment to a technology shock is almost twice as large; see Campbell (1994, Table 3). Then a favorable impulse to technology sets off the process that is the signature of the real business cycle model – a burst of extra employment and a decline in consumption resulting in vigorous capital accumulation.

In addition to the simple assertion of elastic labor supply, the literature proposing fluctuations theories based on that hypothesis has offered three supporting ideas. First was Kydland and Prescott's (1982) use of non-time-separable utility. Second was Rogerson's (1988) observation that workers facing a binary choice between not working and working full time may behave as if they had linear utility and perfectly elastic labor supply. Third was Benhabib, Rogerson and Wright's (1991) consideration of substitution between work in the market and work at home. Their paper marked the beginning of the investigation of margins other than labor–leisure within DSGE models.

A convenient family of non-time-separable preferences follows suggestions of Sargent (1979, p. 371) and Kydland and Prescott (1982) [my discussion is taken from Hall (1991b)]. Let z_t be the accumulated stock of current and past work effort, with persistence factor ω:

$$z_t = (1 - \omega) \sum_{s=0}^{t} \omega^s n_{t-s}. \tag{3.1}$$

My derivation will take n_t to be weeks of work in period t and will assume (realistically) that variations in hours of work per week are small. The parameter ω controls the memory of past work and leisure. If ω is 0, there is no memory; only current work effort matters. If ω is large (close to its upper limit of 1), then z_t depends on a long distributed lag of past work effort. As in my earlier discussion of the baseline

model, the worker orders work schedules with a utility function that is separable over time in the cumulation variable, z:

$$-\sum_{t=0}^{T} \frac{(\bar{n}-z_t)^{1/\sigma-1}}{1/\sigma-1}. \tag{3.2}$$

Define effective leisure as $\bar{n}-z_t$ and actual leisure as $\bar{n}-n_t$. The parameter σ is both the intertemporal elasticity of substitution in effective leisure and the long-run elasticity of substitution in actual leisure (where the long run is enough time so that the distributed lag feature does not matter). In the short run, the elasticity of substitution in actual leisure is greater than σ by an amount that is controlled by the memory parameter, ω. The parameter $\bar{n}$ is the number of weeks physically available for work.

A worker with a high σ will suffer little from a work schedule involving many weeks of work per year in one decade and few weeks per year in another decade, in comparison to putting in the same number of lifetime weeks with no variation from decade to decade. In a situation with free choice of weeks, such a worker will concentrate weeks disproportionately during the years of highest wages.

On the other hand, a worker with low intertemporal substitution (low σ) but high memory persistence, ω (that is, close to one), will tolerate short-term fluctuations in weeks of work but resist decade-to decade movements. Kydland and Prescott use preferences that are slightly more general – current work can have a role in the utility function beyond the role implicit in the variable z_t.

To illustrate the difference between the short-run and medium-run responses of labor supply to wage changes, consider the following question: let $2N$ be the number of periods considered to define the medium run, which might be 24 months. Suppose a worker increases weeks of work by 1 percent in periods $t-N,\ldots,t,\ldots,t+N$. By what percent does the supply price of a week of work in period t increase? The elasticity of labor supply over the $2N+1$-period run is the ratio of the two numbers.

It is convenient to use the λ-constant or Frisch labor supply schedule to answer this question. Let λ be the Lagrangian multiplier associated with the worker's intertemporal budget constraint. The first-order condition associated with labor supply is

$$\frac{\partial U(n_1,\ldots,n_t,\ldots,n_T)}{\partial n_t} = \lambda w_t. \tag{3.3}$$

Here w_t is the real wage in period t stated in period-0 prices, that is, in prices discounted to period 0. The Frisch inverse labor supply function is simply the marginal disamenity of work stated in wage units:

$$\frac{1}{\lambda}\frac{\partial U(n_1,\ldots,n_t,\ldots,n_T)}{\partial n_t}. \tag{3.4}$$

When U is additively separable in labor, this can be solved to give current labor supply as a function of the current wage. Absent separability, it states the supply price

of work in one period as a function of the level of work in that and other periods. Keeping λ constant has two interpretations. First, Equation (3.4) gives the supply price of labor at different points in time along the same labor-supply trajectory. Under this interpretation, statements about the response of the supply price to different levels of work are comparisons of the supply price at different points in time; the change in the level of work is fully anticipated. Second, the supply price conditional on λ has a comparative statics interpretation when the change has little or no effect on λ. Under this interpretation, Equation (3.4) is very similar to (but not quite the same as) the compensated labor supply schedule.

The Frisch labor supply function associated with the preferences considered here is

$$w_t = \frac{1}{\lambda}(1-\omega)\sum_{s=t}^{T}\omega^{s-t}(\bar{n}-z_s)^{-1/\sigma}. \tag{3.5}$$

Let x be the common increment to $n_{t-N},\ldots,n_t,\ldots,n_{t+N}$. For simplicity, assume that the horizon, T, is infinite and that n_t and z_t have the common value n in all periods. Then some manipulations show that the slope of the inverse labor supply schedule is

$$\frac{dw}{dx} = \frac{1}{\lambda\sigma}(\bar{n}-n)^{-1/\sigma-1}\left(1-\frac{2\omega^{N+1}}{1+\omega}\right). \tag{3.6}$$

The elasticity, $\varepsilon(N)$, of the labor supply schedule is

$$\varepsilon(N) = \sigma\frac{\bar{n}-n}{n}\left(1-\frac{2\omega^{N+1}}{1+\omega}\right). \tag{3.7}$$

If there is no memory of past work ($\omega = 0$) or if the displacement of the work is lengthy (N is large), then the elasticity is just the intertemporal elasticity of substitution in leisure, σ, multiplied by the ratio of non-work time to work time:

$$\varepsilon(\infty) = \sigma\frac{\bar{n}-n}{n}. \tag{3.8}$$

The elasticity $\varepsilon(\infty)$ controls labor supply over the life cycle. A worker with an $\varepsilon(\infty)$ of 1 will work twice as many weeks at age 40 as at age 20 if the wage at age 40 is double its level at age 20 (and this doubling was known to be in the offing at age 20). Life-cycle variations in weeks of work do not show an elasticity anywhere near 1 – the evidence appears to favor values of 0.1 to 0.2. If $\frac{\bar{n}-n}{n}$ is $5/47 = 0.11$ and the medium-run elasticity of labor supply is 0.15, then σ is $0.15/0.11 = 1.4$. Here I am considering anticipated life-cycle changes in the wage or changes of short enough duration that feedback through λ can be neglected.

By contrast, the elasticity of labor supply in the context of a one-period displacement ($N = 0$) is

$$\varepsilon(0) = \sigma \frac{\bar{n} - n}{n} \frac{1 + \omega}{1 - \omega}. \tag{3.9}$$

If memory decays at a rate of 20 percent per period, as might be appropriate in a quarterly model, the very-short-run elasticity is 1.8/0.2 = 9 times as large as the medium-run elasticity. The specification is successful in delivering a high short-run elasticity of labor supply without relying on significant decade-to-decade elasticity of labor supply.

3.2. Empirical research

There is a huge empirical literature on labor supply. Pencavel (1986) and Killingsworth and Heckman (1986) survey the direct evidence from panel data on individuals. MaCurdy (1981) is one of the leading studies they consider. The basic approach is the following: the elasticity of labor supply is the ratio of the change in work effort to the change change in wage that occurs as the result of a change in one or another instrumental variable. For example, as a worker moves from age 29 to age 30, the worker's wage typically rises more than normal because this is a steep part of the age-wage-rate profile. If hours of work also rise by more than trend from age 29 to age 30, there is a positive elasticity of intertemporal substitution. Pencavel's (1986) Table 1.22, p. 85, suggests that the elasticity estimated on this basis is somewhere between 0 and 0.45 for men. Card (1994) reviews subsequent research and concludes that Pencavel's earlier conclusion survives unaltered.

Pencavel appropriately devotes considerable attention to the one group of short-run exogenous events whose labor supply effects are well documented – the negative income tax experiments. Experimental subjects experienced a three-year reduction of 30, 50, or 70 percent in effective wages. Pencavel's survey reaches the conclusion that the elasticity of the response of labor supply to these wage reductions was in the fairly narrow range from 0.06 to 0.19 (Table 1.21, p. 80). This finding is the single most telling evidence against the view that intertemporal substitution is high in the short run.

The negative income tax findings are less than definitive for the following reason. For good reasons relating to asymmetric information, workers delegate to their employers the determination of weeks of work. Workers shop among employers with different policies for setting weeks of work, but once the worker accepts a job, the weeks of work required on that job are largely out of the worker's control. In particular, if an event occurs that is personal to the worker, but not within the class of events (such as disability) contemplated by the employment arrangement, it is unlikely that the employer will agree to a reduction in weeks *ad hoc*. Employment arrangements with given, understood rules help control opportunistic behavior by both employers

Table 2
Seasonal averaged deviations from trend[a]

	Quarter			
	1	2	3	4
Number of workers employed	−1.45	0.07	0.30	1.08
Weekly hours	−0.87	−0.14	0.81	0.20
Total hours	−2.32	−0.07	1.11	1.28

[a] Data from Barsky and Miron (1989, Table 2).

and workers. Workers could take extra weeks off by quitting one job and delaying taking another job, but that step dissipates the value of job-specific capital. The finding of small reductions in weeks of work in the negative income tax experiments is not inconsistent with the hypothesis that much larger reductions can occur when the marginal revenue product of labor declines in a downturn. One is unprecedented and unfamiliar, completely new to the environment under which employment arrangements have evolved; the other is exactly within the historical experience that shaped those arrangements.

Another reason that panel studies, both survey and experimental, are not good evidence against elastic short-run supply is the amount of variability they reveal in annual work effort. According to MaCurdy (1981), the standard deviation of annual hours of work around the predictions of his labor supply function is several hundred hours, a significant fraction of the normal level of around 2000 hours. Most of this noise is variation over time around a worker's own normal level of work. If the intertemporal elasticity of labor supply is as low as the numbers in Pencavel's survey, with respect to substitution between one year and the next, then the deadweight burden of the unexplained variability of work is extremely high. A more reasonable conclusion is that the low elasticities apply to life-cycle influences but that much higher elasticities operate at year-to-year frequencies.

A low intertemporal elasticity of substitution in the short run should also make workers averse to predictable seasonal variations in their volume of work. Table 2, taken from Barsky and Miron (1989, Table 2), presents the seasonal averages in percent deviations from trend by quarter found for the private non-agricultural sector of the US economy. The United States has a recession every winter comparable to business-cycle recessions. There is a boom in the summer and fall. Although one might suppose that part of the seasonal movements in hours of work reflects seasonal variations in preferences for work and leisure, it is hard to see how that would result in more work in the summer, during the vacation season, and less work in the winter. If workers had a strong aversion to uneven work schedules, institutions would develop to smooth employment over the season. The seasonal data suggest reasonable amounts of intertemporal substitution among the quarters of the year.

Mulligan (1995) surveys many studies of intertemporal substitution in labor supply, including some not considered in earlier surveys. He stresses novel tests, such as those occurring during exceptional events including wars, the construction of the Alaskan pipeline, and the *Exxon Valdez* cleanup. Unfortunately, he does not consider the negative income tax experiments explicitly but appears to believe that their results should be lumped with panel studies where there are no large exogenous changes in wages (the experiments are mentioned only in footnote 2 and in the summary of his paper). Evidence from the premiums needed to induce brief periods of extraordinary effort or acceptance of irregular work schedules yields higher values of the IES. Mulligan presents evidence on wage premiums in Alaska from the periods of high activity associated with the building of the Alaska pipeline and the *Valdez* cleanup; these range from 0.5 to 2.8. This type of evidence is not fully convincing, however, because if there is a distribution of the IES across workers, only those with the highest values will contribute observations. Mulligan makes an attempt to infer the elasticity of substitution from the experience of World War II, but, as he points out, this is a questionable exercise because, by most measures, real wages were lower during the war than at other times.

Recall that the intertemporal elasticity of labor supply in the baseline neoclassical model is around 4, well above even most of Mulligan's findings. Much higher values than 4 are needed to achieve the degree of magnification required in a realistic macroeconomic model. Intertemporal substitution in labor supply can only be part of the story. Models that do not consider unemployment as a use of time alternative to work appear to be incapable of explaining observed employment volatility along neoclassical lines. Even with an infinite intertemporal elasticity of substitution, as Campbell (1994) shows, the elasticity of employment with respect to the technology shock is still below 1 unless the shock is highly transitory.

3.3. *Unemployment*

As I noted earlier, the most conspicuous shortcoming of the baseline model is its failure to understand unemployment. The mechanism by which workers lose jobs in response to adverse shocks is a promising area to find amplification, and the slow process of re-employment is surely part of the story of persistent periods of slack.

The baseline neoclassical model fails to deal with unemployment in two ways. First, it assumes that the labor market clears instantaneously. Even if workers are leaving some jobs and taking others, the process takes no resources and no time. Second, the model recognizes no heterogeneity in workers or jobs. The model contains no ingredients that would suggest that workers should change jobs – that a worker is more productive in a new job than in the current one. Unemployment cannot be grafted on to the baseline model. A new model, unfortunately much more complicated, is needed to deal effectively with unemployment. New work on job destruction, job creation, and job search has made important advances in this area. Only recently have these ideas been incorporated in DSGE models. Newly developed models achieve employment

Table 3
Alternative measures of quarterly rates of job loss[a]

Source	Quarterly rate of job loss (%)
Permanent separations, UI system data	17.23
CPS tenure survey, 1981	10.04
All separations, Current Population Survey	8.29
Gross employment reductions, LRD	5.66
Permanent layoffs, PSID, 1985	1.81
Displaced workers survey, all workers, 1991–1993	0.61
Displaced workers survey, workers on the job for at least 3 years, 1991–1993	0.59

[a] For sources, see Hall (1995).

amplification by operating on the employment – unemployment margin rather than on the employment – leisure margin. In a sense, they consider labor supply to be elastic because events can cause significant movements of workers between work and job search, even if they cause little movement along the market–nonmarket margin.

Even in normal times, rates of job loss are astonishingly high in the US economy and in other economies like it. Hall (1995) presents data on a number of measures of separation rates. First, the most comprehensive measure of job separations comes from the unemployment insurance system, which includes even the briefest jobs lasting a day or two. Over a quarter, the ratio of total separations to employment is about 17 percent. The same worker can contribute many separations in the same quarter, if engaged in day work, construction, or other high-turnover activities. Second, the tenure survey in the Current Population Survey asks workers how long ago they began their current jobs. Because the shortest category considered is 6 months, this measure does not include multiple separations within a 6-month period, and so, when stated as a quarterly rate, tenure is lower but still a substantial 10 percent per quarter. Third, the Current Population Survey measures separations implicitly, when the same person is reported as working in one month and working at a different job or not working the following month. The rate shown in Table 3 is adjusted for an upward bias that results from random errors in reporting labor-market status in the survey.

The fourth line in Table 3 reports the rate of gross employment reduction, from Davis, Haltiwanger and Schuh (1996). This rate is based on quarterly reductions in employment at the plant level. Because firms sometimes hire during a quarter when total employment falls, the change in employment understates the total separation rate. Their measure is properly called the job destruction rate rather than the separation rate.

The remaining three lines in the table consider separations initiated by employers where the workers had held long-term jobs. These rates are much lower – around one

percent per quarter. A large fraction of total job separations arise from temporary or short-term work.

Rates of job loss rise dramatically at the onset of a recession. Davis and Haltiwanger's rate of job destruction in manufacturing reached peaks of 11 percent and 9 percent per quarter in the recessions of 1975 and 1982, from an average level of 5.7 percent [Davis, Haltiwanger and Schuh (1996)].

The magnitude of job loss when an adverse shock hits the economy is puzzling in some respects. The great majority of workers have been on the job for 5 years or more and expect to remain in the same job for many more years [Hall (1982)]. Higher-tenure workers may have accumulated substantial amounts of job-specific capital, measured as the difference in the expected present discounted value of earnings at their current jobs and the values conditional on departing. Evidence in Ruhm (1991) discussed in Hall (1995) suggests that the typical layoff of a high-tenure worker costs the victim about 1.2 years of earnings, in the form of multiple spells of unemployment and reduced hourly wages. If the anticipated value of job-specific capital is divided evenly between worker and employer, then the typical level of the capital is 2.4 years of earnings, or around $100 000.

It should take a substantial adverse shock to merit the dissipation of $100 000 of specific capital. Labor-market institutions should evolve to protect specific capital against shocks of all kinds, including aggregate ones. Of course, not every job has the typical amount of match capital. Workers with low tenure or in failing businesses may be close to the point where separation would be efficient – it would not lower the joint value of employers and worker. But the evidence at least creates the suspicion that many of the workers who lose their jobs in a recession do not fall into this category. As a general matter, it appears that firms tend to lay workers off despite opportunities to preserve still-valuable job-specific capital. A number of authors have taken this hypothesis as a point of departure for theories of amplification. In addition, recent work has considered the role of heterogeneity in the values of job matches – separations are most likely in the matches whose values are in the lower tail of the distribution, thanks to idiosyncratic factors.

The basics of the theory of job termination are well developed in labor economics[2]. A core question is the efficiency of terminations – efficiency, as usual, means the maximization of joint value. Figure 1 displays the analysis of efficient terminations. The horizontal axis shows earnings available from the next best job in the open market, net of search costs. The vertical axis shows the worker's marginal product at this employer. Separation should occur below the 45° line. Whether the separation is initiated by the worker as a quit or by the employer as a layoff depends on the details of the employment arrangement[3].

[2] For example, Hashimoto and Yu (1980), Hall and Lazear (1984), and McLaughlin (1991).

[3] See McLaughlin (1991).

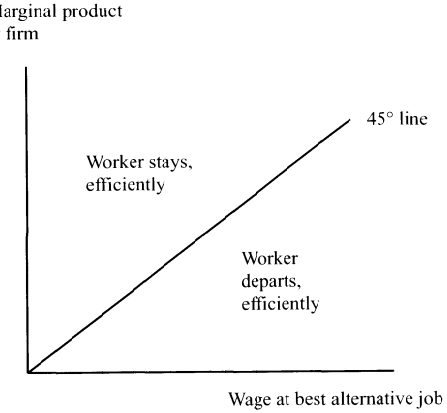


Fig. 1. Efficient separation.

Efficient separations would be likely if the variables in Figure 1 were observed by the employer and the worker. For example, if the worker can locate the best possible outside wage offer costlessly, and the employer can verify the offer, then the employer will match the offer and retain the worker if the offer is below the worker's marginal product, and let the worker accept the offer otherwise. This arrangement does not require the marginal product to be observable[4].

When neither party to the employment contract can verify the other's data, efficiency is more of a challenge. Any provision granting the employer the right to lower compensation after a worker has accepted employment and made job-specific investments will invite opportunistic wage cuts. Even when demand has truly fallen, and renegotiation of the terms of employment is appropriate to retain the worker efficiently, the worker will not be able to verify that the employer is not trying to deprive the worker of job-specific rents. Suppression of renegotiation may be an important feature of employment arrangements.

Absent government prohibition of certain types of governance, one would expect the form of the employment relationship to evolve to maximize the joint value achieved by employers and workers. In principle, this proposition should apply even if either the employer or the employee, or both, have market power. Maximization of joint value will occur subject to the constraints of limited abilities to observe or verify key measures and the likelihood that many workers are unable to borrow as much as they would otherwise against future earnings.

Most jobs have specific capital. Workers develop skills related to the employer's particular way of doing business. They develop personal relationships with their co-workers. They may choose places to live and particular houses based on their

[4] Hall and Lilien (1979) discussed efficient employment arrangements with unilateral information private to employers.

employer's location. Firms accumulate valuable knowledge about their workers' skills. More subtle employment practices may be needed to protect investments in specific capital.

When the efficiency of the continuation of the match is a live issue, protection of specific investments becomes a serious challenge. In that case, some kind of joint or unilateral procedure is needed to determine if a match should continue or end. If either party has the power to end the job (the worker to quit or the employer to terminate), the party can use that power to deprive the other party of the expected return to the investment. For example, an employer might attract a worker to make an expensive move by offering a high salary. A year later, the employer might approach the worker and say that the worker would be terminated unless the worker accepted a much lower salary. The worker would accept the salary reduction as long as the salary remained above the value of the next best job, which might involve another expensive move. An employment arrangement can include severance pay to limit this type of opportunistic behavior by employers.

3.3.1. Mechanism design and labor contracts

The discussion of the design of the labor contract has been strongly influenced by the literature on mechanism design derived from Mirrlees's (1971) famous paper. A key idea in this literature is that contracts can only be contingent on measures that are verifiable – it is not enough that the measures be observable. Hart (1983) discusses the first round of thinking along these lines, where separation or other employment decisions are made unilaterally by worker or firm, subject to a contract determined in advance. A more recent elaboration of the theory of the employment relationship in the mechanism design framework is in the work of Charles Kahn and Gur Huberman (1988). In their model, the worker's productivity is observed only by the employer, but the productivity depends on an investment in specific capital observed only by the worker. Absent both of these information limitations, simple contracts would give the first-best outcome. If productivity were verifiable, then the wage would be contingent on actual productivity, and the worker would have the right incentive to make the investment. If the investment were itself observable, the employer would reward the worker for making the investment. With both unobservable, the following more complicated contract delivers the efficient outcome: The parties agree in advance on a wage to be paid after the investment is made. Upon observing the worker's productivity later, the employer can either keep the worker and pay the wage, or discharge the worker. The worker does in fact make the investment and is retained, which is the efficient outcome.

Gilson and Mnookin (1990) argue that the up or out rule common in law firms is the result of suppression of renegotiation. In order to induce associates to make firm-specific investments, the firm promises not to offer the associate a salary just above the best outside salary. Instead, at a predetermined time, the firm chooses between offering partnership or terminating the associate.

Although Kahn and Huberman do not stress the point, suppression of renegotiation is central to the success of their contract. After the worker has made the firm-specific investment, the employer could say, "If I have to pay you the wage we agreed upon, I won't keep you. But if you agree to a lower wage, I will keep you." There is no violation of the contract in this offer. But if the worker anticipates that the employer is free to make this offer, the worker will not make the investment and the scheme will fail.

Considered as a game played only once, the Kahn–Huberman contract fails the test of credibility – it is not subgame perfect. Suppression of renegotiation requires the employer to commit not to take a step later that would be rational and permitted under the terms of the contract. The problem is the same as the one studied extensively by monetary economists (a central bank needs some way to commit not to create a monetary surprise later, when such a surprise would be rational later) and in public finance (the tax authorities need some way to commit not to levy a capital tax later, when such a tax is the ideal, neutral lump-sum tax later)[5].

Although in both the monetary and fiscal settings, there are no formal institutions to enforce commitments, experience – especially recently – suggests that something like the favorable equilibrium with commitment can be achieved anyway. Reputations of policymakers and institutions seem to be an important part of the story – see Barro and Gordon (1983) in the analogous context of monetary policy. Reputation may explain the credibility of the suppression of renegotiation as well. If an employer is expected to remain in business permanently, it will pay for it to develop a reputation for adhering to policies of not renegotiating. The concept of reputation can be explained in models of games of repeated play, or in other frameworks[6].

Suppression of renegotiation seems to be an important part of the cultural norms of the labor market as well. The offer to retain an employee by departing from previously announced standards of compensation is seen as morally wrong. Standards of ethical conduct support up-or-out rules in universities and professional practices. It is wrong to extend a non-tenured faculty member's appointment after denial of tenure, even though both sides favor it.

Truman Bewley's extensive field study of employment relationships in a depressed local labor market documents the absence of renegotiation[7]. By far the most common reason given by employers and their advisers for not rewriting employment arrangements in order to preserve jobs is that lowering wages would destroy morale. In other words, workers see a departure from the established compensation patterns as a violation of the rules of the workplace. They think it is wrong to depart from the principle that employers unwilling to pay promised levels of compensation should discharge their workers.

[5] See Fischer (1980). On the general issue of the value of commitment in games, see Fudenberg and Tirole (1991, pp. 74–77).

[6] See Carmichael (1984).

[7] Bewley (1994).

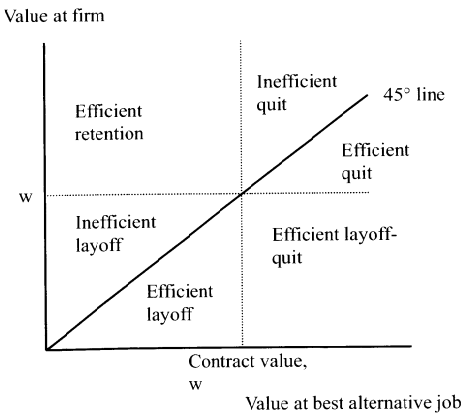


Fig. 2. Wage contract with suppression of renegotiation.

Suppression of renegotiation has some of the implications of the types of wage rigidity considered in macroeconomics. But it does not explain any failure of the labor market to clear. It puts no restriction on the terms under which new workers are hired; it is completely consistent with market clearing in the market for new hires. Figure 2 shows that suppression of renegotiation results in excess, inefficient separations. Suppose that the worker was hired with the understanding that the wage would be w. The firm has the right to terminate the worker if the wage falls below w and the worker has the right to quit if there is an alternative job paying more than w. The standard for efficiency remains as in Figure 1 and does not involve the contract wage. Figure 2 shows that a separation will always occur if the match has become inefficient. In the inefficient area below the 45° line, either a quit will occur (the triangle on the right) or a layoff will occur (the triangle at the lower left), or both a quit and a layoff will occur (the lower right quadrant). The suppression of renegotiation permits the destruction of efficient matches, however. When conditions are good in the outside market, relative to the contract wage, but conditions are even better at this employer (the triangle at the upper right), the worker quits even though the two parties could renegotiate to mutual advantage – there is a wage that will keep employment profitable for the employer but also exceed the worker's best alternative wage.

At the left in Figure 2 is the case of greatest interest in this chapter. An inefficient layoff occurs when conditions are bad at the firm but even worse in the outside market. The parties fail to renegotiate a wage reduction, even though a mutually beneficial one is available.

Finally, the upper left quadrant describes a success for the contract; retention is efficient and it actually happens.

3.3.2. *The modern strategic view of the employment relationship*

Modern thinking about the employment relationship, as reflected in Diamond (1982a, b), Mortensen (1982), Ramey and Watson (1997), Caballero and Hammour (1996, 1998), and many other papers, insists that strategic relationships, such as the employment relationship, satisfy the criterion of subgame perfection. Parties will not adhere to terms such as fixed wages that they can negotiate around later to their mutual benefit. An employment contract does not necessarily guide the relationship as stated, but only establishes the threat points for a subsequent bilateral bargaining problem. The Nash solution to that bargaining problem – where the parties split the joint surplus from their relationship – governs the outcome. For further discussion of the relation between the two branches, see Aghion, Dewatripont and Rey (1990) and Hall (1995, 1997)[8].

3.3.3. *Efficiency wages*

The theory of efficiency wages – Shapiro and Stiglitz (1984) and Akerlof and Yellen (1990) – has had an important role in macroeconomics for more than a decade. At a minimum, the theory helps us understand why the natural or chronic level of unemployment is so high. Employers seek to create idiosyncratic value for their workers in their jobs, so that the threat to terminate deters misconduct such as shirking. Absent unemployment, employers will pay wage premiums to create the needed idiosyncratic value. Because it is impossible for every employer to pay a premium, there is unemployment in equilibrium. Note that unemployment from efficiency-wage factors is not a socially or privately valuable use of time directly – efficiency-wage models do not suggest that workers have elastic labor supply in the conventional sense.

A number of writers have discussed the ways that efficiency wages contribute to amplification – see, for example, Danthine and Donaldson (1995), Picard (1993, Chapter 7), Phelps (1994) and D. Romer (1996, p. 220). Woodford's (1994b) discussion of Phelps reveals the benefits of a full DSGE treatment, as some of Phelps's proposed mechanisms may not work as he describes, once all feedbacks are considered. Woodford (1994a) shows that technology shocks have no more effect in a DSGE model with efficiency wages than in an otherwise similar one with a neoclassical labor market. MacLeod, Malcomson and Gomme (1994) develop a DSGE model where unemployment arises only because of efficiency wages. The model is able to match the volatility of output but falls short of matching the volatility of employment. The authors suggest that adding elastic labor supply as a second amplification mechanism would overcome this problem.

[8] The recent literature on contracts in general has gone through a similar transformation – see, for example, Segal and Whinston (1996, 1997).

3.3.4. Job destruction

Models with dynamic labor markets that incorporate job destruction, job search, and job creation have made important progress in explaining amplification. Job loss after an adverse shock seems a natural way to model employment reductions. As Mortensen and Pissarides (1994) point out, it is natural to assume the absence of adjustment costs for job destruction. As a result, there is an important asymmetry in the adjustment of employment. The level of employment can decline immediately when an adverse shock strikes, whereas the rate of increase is limited by the costs of hiring, which are convex in the hiring rate. Heterogeneity is essential to the story – if workers and jobs are homogeneous, there is no reason for matches to break up and no time required to put workers back in jobs.

Wright (1986) appears the first to build a modern general equilibrium model where unemployment arises from heterogeneity and time-consuming matching. In Wright's model, as in a number of successors, separation is an exogenous event. Workers are thoughtful about selecting a new job from the heterogeneous set that is available. Search is time-consuming because workers find out about only one new job per period of search. Aggregate unemployment can jump upward because of an information limitation that can cause all searchers to make the inference that it is a good idea to wait for a better job to come along [the information setup is similar to the one in Lucas (1972)].

Mortensen and Pissarides (1994) is a leading example of current thought about job destruction and unemployment, although theirs is not a full DSGE model – product prices are taken as exogenous. The authors avoid dealing with a full model of the heterogeneity of job–worker matches through the assumption that the productivity of an existing match is drawn from the same distribution for all matches, not from a distribution that depends on the history of the match. Both job destruction and job creation are the result of rational economic behavior by employers and workers. Job destruction occurs at the moment that the joint surplus from the match passes through zero. A key concept in the theory is the option value of the employment relationship, which is an element of joint value.

Merz (1995) and Andolfatto (1996) developed DSGE models with job search. To remain within the representative agent general-equilibrium framework – with huge resulting simplification – they presume that families or broader institutions insure individual workers against job loss. In these papers, jobs end at random, not in response to economic forces. But the levels of unemployment and employment fluctuate because of changes in the rate of job creation driven by aggregate conditions. These models achieve amplification, as discussed earlier, by opening up the employment-search margin. Because they do not permit jumps in job destruction along the lines suggested by Mortensen and Pissarides, the amount of amplification is limited.

Caballero and Hammour (1996, 1998) have developed a model of endogenous job destruction. Caballero, Engel and Haltiwanger (1997) apply the ideas of that work empirically. A job is destroyed the moment that its match value crosses zero,

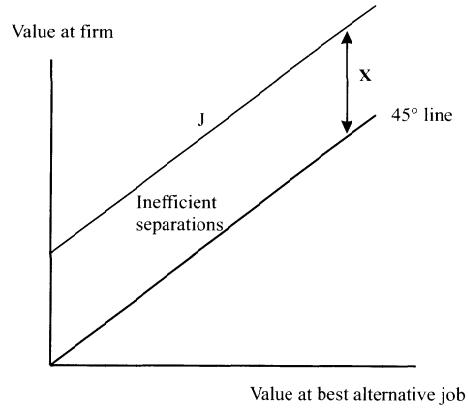


Fig. 3. Separations in the Ramey–Watson model.

in accord with the efficiency condition discussed earlier. The value of each match begins at a positive level, reflecting match capital created by the worker's search effort and the employer's recruiting effort. Thereafter, it evolves as idiosyncratic and aggregate shocks perturb product demand, productivity, and the worker's alternative opportunities. In their general form, this type of model is hard to handle because the state of the economy at each moment includes the distribution of matches by current value, which has a dimension equal to the number of matches.

Gomes, Greenwood and Rebelo (1997) tackle the hard problem of studying a DSGE model without the simplifying assumptions of earlier authors. Their workers do not enjoy the perfect unemployment insurance assumed in earlier models. The value of a match evolves according to a stochastic process with memory, where every match is also influenced by an aggregate variable, so the distribution of workers by current match productivity cannot be simplified.

Efficiency wages have been brought back into the picture in the DSGE framework by G. Ramey and Watson (1997). Figure 3 strips their model down to its bare essentials. Workers can enjoy a benefit, X, if they misbehave – for example, X might be the amount they could steal. Misbehavior is detected with certainty by the employer, but cannot be proven in court, so it cannot be a contingency in a contract. Unless workers have a personal value from continuing in the job of at least X, they will take X and then find another job.

The zone of inefficient separations in the Ramey–Watson model is the area between the 45° line and the line that is X above the 45° line. A job could have substantial joint value, say at J in the figure. But a small shock could move that job below the upper line, causing a separation. The employment relationship is unnecessarily fragile (compared to its full-information version) because it breaks up whenever the joint value achieved by the match falls below X, rather than surviving unless it falls to zero.

In what sense is there amplification in the Ramey–Watson model? Jobs are at risk to small shocks even though they have positive amounts of joint job-specific

value. In a world of homogeneous worker–employer matches, their model explains job destruction without requiring that all matches have little match value. On the other hand, with heterogeneous match values, as in Caballero and Hammour (1996) and related research, the contribution of efficiency wages to amplification is less clear. In those models, job destruction occurs when match values drift down to zero, under the influence of random idiosyncratic and aggregate influences. The Ramey–Watson setup changes the boundary point where job destruction occurs, but does not change the magnitude of the response to adverse aggregate shocks in an obvious way.

3.3.5. *Reorganization and reallocation*

Lilien (1982) began the literature suggesting that reorganization or reallocation could create aggregate effects on employment and unemployment. Impulses that have no net effect on aggregate productivity, for example, nonetheless raise unemployment during the period when they cause workers to move from one sector to another. Reorganization is an activity that is a close substitute for production [Hall (1991a)]. The most conspicuous form of reorganization is the movement of workers from unsuccessful productive units to new or growing units. Other forms of reorganization include relocation and re-employment of capital and the rearrangement of contractual relations among units. One can think of these activities as investment flows that form organizational capital. As in the case of the use of output either for consumption or capital formation, production and organizational capital formation are perfect substitutes. Perfect substitution is a natural source of amplification, just as is the perfect substitution between market work and time spent in other activities.

Flows into reorganization occur out of jobs where the match value has just reached zero, possibly as the result of an aggregate impulse. Match value will fall either if workers' marginal contributions fall at their current jobs or if the likely value of employment elsewhere rises, net of search costs. Thus three types of impulses are amplified in the reorganizational view:

(1) Impulses that raise workers' contributions in some sectors and lower them in others, without affecting the efficiency of search, such as changes in the composition of product demand.
(2) Impulses that lower productivity in all sectors but do not affect the efficiency of search, such as reductions in productivity in all sectors.
(3) Impulses that do not affect productivity but raise the efficiency of search, such as streamlining the labor market.

Data on cyclical flows of labor are helpful in understanding reorganization. Flows out of various industries – measured either as gross job destruction [Davis, Haltiwanger and Schuh (1996)] or as net employment change – are highly correlated across industries. In particular, the abrupt shedding of workers during a recession occurs in most industries. Impulses that stimulate movements of workers from one sector to another cannot explain the most important facts about recessions.

An exogenous improvement in the efficiency of job search relative to the productivity of employment would explain the burst of job separation and high volume of job-seeking in recessions. This explanation is a close relative to the one based on productivity at home rising relative to productivity in the market – workers are sucked out of employment because of the rising attractiveness of an alternative use of time. It shares the defects of that view discussed earlier. It does have the advantage of explaining – however implausibly – the sharp increase in unemployment in recessions, a phenomenon not considered in the existing home production models.

Endogenous changes in job search costs are a more promising way to make reorganization a believable element in fluctuations theory. They arise in models of complementarity. Peter Diamond (1982b) began modern rigorous thinking on this topic. In his model, producer–consumers search for each other. Upon making a match, each can consume and then produce again. The presence of one search agent confers an external benefit on others by raising the probability of encountering a partner. This is a thick-market externality. In Diamond's model, two or more equilibria are possible. In one there are relatively few searchers ready to trade. Because it is difficult to trade, less production takes place. It is a decentralized equilibrium. In the superior equilibrium, more traders are in the market, so more production takes place. Again, this is a decentralized equilibrium.

Diamond's model – like many successors based on complementarities – contains the ultimate form of amplification, indeterminacy. The tiniest impulse, including sunspots, could trigger a move from one equilibrium to another, with very different levels of output and employment.

As stressed in a number of places in this chapter, realistic fluctuations models need to consider unemployment seriously and explicitly. Most research on complementarities has concentrated on output rather than unemployment. Hall (1991a) makes an attempt to apply the logic of Diamond's model of search complementarities to unemployment over the business cycle. The paper offers a crude measure of endogenous increases in search efficiency when unemployment is higher.

At one level, it must be true that it is better for one worker to search when many others are searching as well. Like most other economic activities, job search is concentrated during the daylight hours Monday through Friday. It is efficient to concentrate job–worker matching during a limited set of hours of the week. Many job markets – including the one for economists – are highly seasonal. Again, temporal concentration of matching activities is efficient. It is a leap from these observations, however, to the conclusion that recessions are good times to look for work in the same way that Tuesday at 3 pm and the first week in January are good times.

Before rejecting the view that there are increasing returns to aggregate search, one should consider carefully the evidence developed by Davis, Haltiwanger and Schuh (1996) that the hiring rate (measured by gross job creation) reaches a startling peak immediately after the spike of job destruction that occurs during the initial contraction phase of a recession. The extended period of high unemployment following

a contraction is a period of matching frenzy, with both job destruction and job creation at abnormally high levels.

Evidence cited by Hall (1991a) on the cyclical behavior of job-finding rates is mixed. Blanchard and Diamond (1990) report that a recession that raises unemployment by two percentage points reduces the job-finding rate from a normal level of 24.0 percent per month to 21.8 percent per month. On the other hand, Hall reports a regression relating the Davis–Haltiwanger measure of the volume of job–worker matching to the level of unemployment. He finds an increasing marginal benefit from the stock of unemployment on the flow of new matches.

Although the amplification mechanism based on endogenous improvements in search efficiency during recessions is on uncertain ground, the evidence just reviewed raises serious doubts about the opposite (and conventional) view that recessions are times when jobs become much harder to find. A reasonable intermediate view is that search efficiency is about the same at high and low unemployment. This disposes of a potential attenuation mechanism – job matches would be more stable in recessions than normal times if job search became more costly in recessions.

4. Persistence

The time-series properties of the principal macro variables are reasonably well understood. Unemployment is stationary – it returns about one third of the way to its normal level each year after a shock displaces it[9]. Output and employment have both cyclical and highly persistent – possibly integrated – components. The persistence mechanisms in a fluctuations model need to be able to explain the stationary but serially correlated movements of unemployment and the corresponding cyclical movements of output and employment. The highly persistent components of employment and output derive from slow-moving changes in preferences and technology and are not in the domain of the persistence mechanism of the fluctuations model.

Although a number of authors have identified sources of persistence other than the mechanics of job search [such as Burnside and Eichenbaum (1996) and Saint-Paul (1996)], I will focus mainly on this single topic, which dominates current thinking about persistence.

4.1. Time-consuming matching in the labor market

One of the most interesting and successful recent developments in the labor side of macroeconomics has been the development of modern models of job search. Diamond (1982b) and Mortensen (1982) are the starting points. Good summaries are

[9] In the USA. In many other countries, unemployment is close to a random walk.

in Mortensen (1986), Pissarides (1990), and Romer (1996, chapter 10). My discussion will be brief because Mortensen's chapter in this volume covers this area in detail.

In the standard matching model, a random meeting occurs between a job seeker and an employer. A match occurs if it increases the joint value of the two parties, in which case they divide the joint surplus. The simplest model has a constant probability that a job-seeker will be matched. The persistence parameter for aggregate unemployment – the serial correlation coefficient – is controlled by the job-finding probability. The matching model provides a simple and elegant persistence mechanism for a general equilibrium macro model.

From the start, it has been clear that it is an uphill battle to use the matching model to explain the actual persistence of unemployment in the USA. To see the relation between the job-matching rate and the serial correlation of unemployment, consider the following elementary model. Let d_t be job destruction, n_t be employment, $\bar{n}$ be the fixed supply of labor, $u_t = \bar{n} - n_t$ be unemployment, and f the per-period job-finding rate. Then employment this period consists of employment last period plus those among the unemployed who found new jobs less the number of jobs destroyed:

$$n_t = n_{t-1} + fu_t - d_t \tag{4.1}$$

or

$$u_t = (1 - f)\, u_{t-1} + d_t. \tag{4.2}$$

Thus, if job destruction is white noise, unemployment follows an AR(1) process with serial correlation $1 - f$.

As I noted earlier, the average job-finding rate is about 24 percent per month. The monthly serial correlation of unemployment is 0.988, which would imply a job-finding rate of only 1.2 percent per month. There is a discrepancy of a factor of 20 between the time-series properties of unemployment and the job-finding rates experienced by individuals. Cole and Rogerson (1996) have studied this discrepancy and concluded, "Our main finding is that the [matching] model can account for the business cycle facts, but only if the average duration of a non-employment spell is relatively high – about nine months or longer." With an average job-finding rate of 24 percent per month, the average duration is actually much less. Something is missing from the simple model.

4.2. The importance of secondary job loss for persistence

Hall (1995) suggests that the missing element is induced secondary job loss. The first job that a recently discharged worker finds may be an explicitly temporary job, or it may turn out to be a bad match once the worker joins the firm, or the match may break soon because, in its early stages, it has little job-specific capital. There is evidence of large amounts of secondary spells of unemployment following an initial impulse.

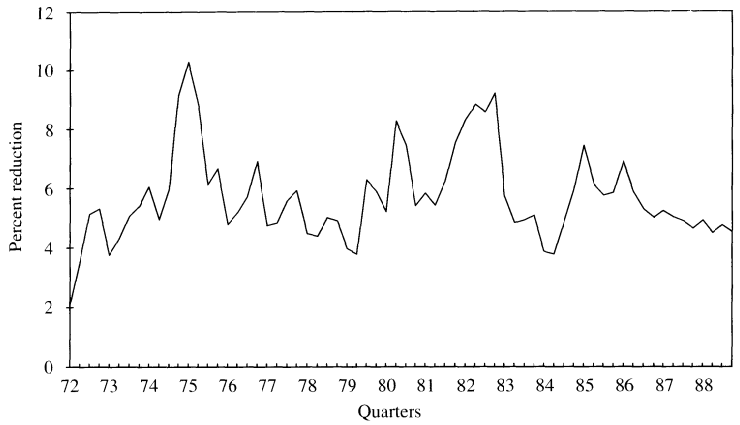


Fig. 4. Gross employment reduction rate in manufacturing.

Figure 4 shows quarterly data on gross employment reductions in manufacturing, taken from the work of Davis, Haltiwanger and Schuh. Gross employment reductions are measured at the level of individual plants. The series shows the total reduction in employment at plants where employment fell from one quarter to the next, as a percent of total employment. Gross employment reductions appear to be the best available measure of the immediate effect of adverse macroeconomic events on the labor market. In particular, as Figure 4 shows, recessions start off with large bursts of employment reductions. The flow of gross employment reductions is not persistent; during the extended slump after a sharp contraction, gross employment reductions are at normal levels[10]. Persistence in unemployment and employment appear to come from other sources.

Data on the flow of workers into unemployment provide another, quite different view of the dynamics of job loss. The best data for this purpose show the flow from permanent layoffs alone, separately from temporary layoffs, quits, new entrants, and re-entrants. Figure 5 shows these data since they became available in 1976[11]. New permanent layoffs are much more persistent than gross job reductions. A burst of job reductions, as in 1982, is followed by several years of higher new permanent layoffs. The data have a strong distributed lag relationship – see Hall (1995).

A number of factors enter the explanation of the lag from employment reductions to new unemployment. First, employment reductions are measured only in manufacturing,

[10] As Davis, Haltiwanger and Schuh note, plant level employment is highly persistent; it is essentially a random walk. Hence the flow of reductions is close to white noise.

[11] The data come from the Current Population Survey and are published in *Employment and Earnings*. They refer to workers who became unemployed as a result of permanent layoff, whose unemployment began within five weeks of the survey.

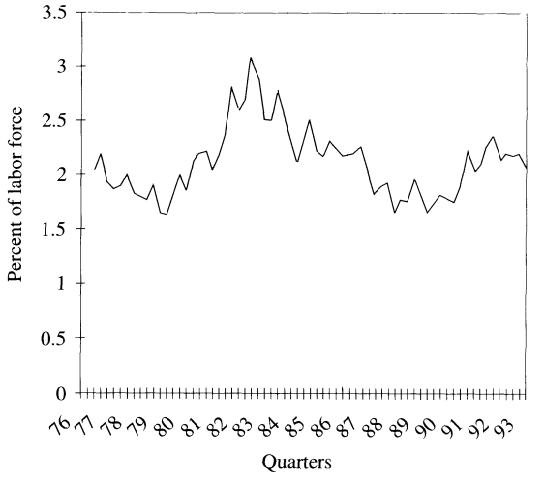


Fig. 5. New permanent layoff unemployment. Source: Bureau of Labor Statistics, Current Population Survey, permanent layoff unemployed less than 5 weeks.

whereas new unemployment is measured economy-wide[12]. A systematic lag of non-manufacturing behind manufacturing would explain some part of the lag. Second, many workers who lose their jobs do not become unemployed – they move immediately to other jobs or leave the labor force. During the period of slack labor markets following a burst of employment reductions, a larger fraction of job-losers become unemployed. Third, permanent job loss has important delayed effects. Many of the workers who move quickly to other jobs have taken temporary work, either jobs with predetermined short terms, or those with naturally high turnover. Those who left the labor force upon loss of a long-term job often re-enter the labor force.

The micro and macro evidence suggests strongly that terminations beget later terminations. When an event breaks a set of long-term employment relationships, the workers released into the labor market will form new relationships. Many of the new jobs will prove to be short-lived. First, it may make sense for an individual to take a temporary job while looking for a new permanent job. Second, a worker long out of the market may experiment with alternative types of work before finding a good long-term match. Third, employers may have explicit policies of hiring many candidates and keeping only the fraction who prove to be well matched. Fourth, immediately after being hired, the typical worker will be close to the margin for discharge, either by the standards of the efficient separation model or the models of suppressed renegotiation or efficiency wages. Both the systematic accumulation of match-specific capital and the random accumulation of rent will have had little time to occur. Low-tenure workers

[12] In principle, data on new permanent-layoff unemployment among workers previously employed in manufacturing could be tabulated from the Current Population Survey, but it would require processing all of the monthly CPS tapes. I do not believe this has yet been done.

are the logical candidates for separation – last hired, first fired is the rational separation rule under broad conditions.

A specific adverse event will create an immediate burst of terminations, followed by the second, third, and subsequent rounds of terminations. Induced subsequent job losses seem to be a promising explanation of persistence. Following a single adverse shock, employment will be depressed and unemployment elevated by subsequent rounds of adjustment in the labor market.

A glance at the data show that a simple model of transitions between jobs and search cannot be faithful to even the most conspicuous features of the market's dynamics. Rates of separation from jobs decline sharply with tenure on the job, and job-finding rates fall with the duration of unemployment. Part of the duration dependence is genuine and part reflects the sorting of heterogeneous workers[13]. Moreover, previous history appears to influence transition rates. For example, workers terminated from long-term jobs have lower job-finding rates than do other searchers, are more likely to lose subsequent jobs than are other short-tenure workers, and have lower job-finding rates in subsequent spells of unemployment.

Some basic properties of job loss have emerged in this review of the evidence. Microeconomic studies of serious job loss show significant downstream effects on the subsequent experiences of individuals in the labor market. Loss of a long-term job leads to a period of episodic employment, periods of job search or time out of the labor market, and lower earnings when working. The effects extend for at least four years. In the macroeconomic evidence, bursts of gross employment reductions coincide with abnormal levels of serious job loss. The downstream effects visible in time series data for unemployment are similar to the effects found in micro data for individuals.

The macro data show occasional sharp disruptions of employment followed by long periods of rebuilding of employment relationships. This rebuilding may be an important part of the propagation mechanism of the business cycle. The length of time that the economy takes to recover from an adverse shock has perplexed macroeconomists for many years. Rebuilding may help solve this puzzle of persistence.

Den Haan, Ramey and Watson (1997) have developed a DSGE model with realistic persistence in which efficient job destruction interacts with capital formation. They provide an alternative explanation of induced secondary job loss. A key property of their model is that the idiosyncratic shock at the level of the plant or individual job match is unpredictable white noise. An aggregate shock results in a first round of job destruction. There follows a period of high interest rates during which the threshold value for the idiosyncratic shock changes so as to increase the probability of job destruction. Until the aggregate shock wears off, job destruction continues at abnormally high levels. The model is successful in explaining the persistence of job destruction and unemployment, without invoking unrealistically low rates of job finding. On the other hand, it relies on highly persistent technology shocks (with a quarterly serial correlation of 0.95) in order to generate persistent changes in interest rates. The model's assumption that the idiosyncratic component of job match value is white noise is also intrinsic to the model's success in explaining persistence. Under the more realistic assumption of a random walk for the idiosyncratic component, all of the job destruction triggered by an aggregate shift in technology would occur immediately and there would be no persistent subsequent job destruction.

[13] See Heckman and Singer (1985) and Devine and Kiefer (1991).

5. Conclusion

In the economies of the USA and other modern countries, large responses, especially recessions, seem to result from small impulses. Their effects on the economy must operate through an amplification mechanism. The fragility of the employment relationship seems to underlie that sensitivity. Despite substantial job-specific capital in the majority of jobs, millions of workers are released into the labor market during each contraction. The resulting unemployment is persistent. Not only does it take time for workers displaced by a recession to find new jobs, but the average one has to find several new jobs, a process that stretches over about four years.

DSGE models have come a long way since Kydland and Prescott (1982) in incorporating labor-market frictions and giving correspondingly more realistic portrayals of the economy. Recognition of the heterogeneity of workers and jobs has been central to this improvement in macro modeling.

Acknowledgments

This research was supported by the National Science Foundation under grant SBR-9410039 and is part of the NBER's research program in Economic Fluctuations and Growth. I am grateful to the editors for helpful comments.

References

Aghion, P., M. Dewatripont and P. Rey (1990), "On renegotiation design", European Economic Review 34:322–329.

Aiyagari, S.R., L.J. Christiano and M. Eichenbaum (1992), "The output, employment, and interest rate effects of government consumption", Journal of Monetary Economics 30:73–86.

Akerlof, G.A., and J.L. Yellen (1990), "The fair wage-effort hypothesis and unemployment", Quarterly Journal of Economics 105:255–283.

Andolfatto, D. (1996), "Business cycles and labor-market search", American Economic Review 86: 112–132.

Barro, R.J., and D.B. Gordon (1983), "Rules, discretion, and reputation in a model of monetary policy", Journal of Monetary Economics 12:101–121.

Barsky, R.B., and J.A. Miron (1989), "The seasonal cycle and the business cycle", Journal of Political Economy 97:503–534.
Basu, S. (1996), "Procyclical productivity: increasing returns or cyclical utilization?", Quarterly Journal of Economics 111:719–751.
Basu, S., and J.G. Fernald (1997), "Returns to scale in U.S. production: estimates and implications", Journal of Political Economy 105:249–283.
Benhabib, J., R. Rogerson and R. Wright (1991), "Homework in macroeconomics: household production and aggregate fluctuations", Journal of Political Economy 99:1166–1187.
Beveridge, S., and C.R. Nelson (1981), "A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to the measurement of the 'business cycle'", Journal of Monetary Economics 7:151–174.
Bewley, T. (1994), A Field Study on Downward Wage Rigidity (Yale University).
Blanchard, O.J., and P.A. Diamond (1990), "The cyclical behavior of the gross flows of U.S. workers", Brookings Papers on Economic Activity 1990(2):85–155.
Burnside, C., and M. Eichenbaum (1996), Factor hoarding and the propagation of business cycle shocks", American Economic Review 86:1154–1174.
Burnside, C., M. Eichenbaum and S.T. Rebelo (1993), "Labor hoarding and the business cycle", Journal of Political Economy 101:245–273.
Caballero, R.J., and M.L. Hammour (1996), "On the timing and efficiency of creative destruction", Quarterly Journal of Economics 111:805–852.
Caballero, R.J., and M.L. Hammour (1998), "Jobless growth: appropriability, factor substitution, and unemployment", Carnegie-Rochester Conference Series on Public Policy 48(June):51–94.
Caballero, R.J., E. Engel and J. Haltiwanger (1997), "Aggregate employment dynamics: building from microeconomic evidence", American Economic Review 87:115–137.
Campbell, J.Y. (1994), "Inspecting the mechanism: an analytical approach to the stochastic growth model", Journal of Monetary Economics 33:463–506.
Card, D. (1994), "Intertemporal labor supply", in: C.A. Sims, ed., Advances in Econometrics: Sixth World Congress of the Econometric Society, vol. II (Cambridge University Press) 49–78.
Carmichael, H.L. (1984), "Reputations in the labor market", American Economic Review 74:713–725.
Cogley, T., and J.M. Nason (1995), "Output dynamics in real-business-cycle models", American Economic Review 85:492–511.
Cole, H.L., and R. Rogerson (1996), "Can the Mortensen–Pissarides matching model match the business cycle facts?", Staff Report 224 (Federal Reserve Bank of Minneapolis, Research Department, December).
Danthine, J.-P., and J.B. Donaldson (1995), "Non-Walrasian economies", in: T.F. Cooley, ed., Frontiers of Business Cycle Research (Princeton University Press, Princeton) 217–242.
Davis, S.J., J.C. Haltiwanger and S. Schuh (1996), Job Creation and Destruction (MIT Press, Cambridge, MA).
den Haan, W.J., G. Ramey and J. Watson (1997), Job Destruction and Propagation of Shocks (University of California, San Diego).
Devine, T.J., and N.M. Kiefer (1991), Empirical Labor Economics: The Search Approach (Oxford University Press, New York).
Diamond, P.A. (1982a), "Wage determination and efficiency in search equilibrium", Review of Economic Studies 29:217–227.
Diamond, P.A. (1982b), "Aggregate demand management in search equilibrium", Journal of Political Economy 90:881–894.
Fischer, S. (1980), "Dynamic inconsistency, cooperation and the benevolent dissembling government", Journal of Economic Dynamics and Control 2:93–107.
Fudenberg, D., and J. Tirole (1991), Game Theory (MIT Press, Cambridge, MA).
Gilson, R.J., and R.H. Mnookin (1990), "The implicit contract for corporate law firm associates: ex post opportunism and ex ante bonding", in: M. Aoki, B. Gustafsson and O. Williamson, eds., The Firm as a Nexus of Treaties (Swedish Collegium for Advanced Study in Social Sciences Series, London) 209–236.
Gomes, J., J. Greenwood and S.T. Rebelo (1997), Equilibrium Unemployment (University of Rochester, Rochester, NY).
Hall, R.E. (1982), "The importance of lifetime jobs in the U.S. economy", American Economic Review 72:716–724.
Hall, R.E. (1988), "The relation between price and marginal cost in U.S. industry", Journal of Political Economy 96:921–947.
Hall, R.E. (1990), "Invariance properties of Solow's productivity residual", in: P. Diamond, ed., Growth/Productivity/Unemployment: Essays to Celebrate Robert Solow's Birthday (MIT Press, Cambridge, MA) 71–112.
Hall, R.E. (1991a), "Labor demand, labor supply, and employment volatility", National Bureau of Economic Research Annual:17–46.
Hall, R.E. (1991b), "Substitution over time in consumption and work", in: L. McKenzie and S. Zamagni, eds., Value and Capital Fifty Years Later (MacMillan) 239–267.
Hall, R.E. (1995), "Lost jobs", Brookings Papers on Economic Activity 1995(1):221–273.
Hall, R.E. (1997), "Macroeconomic fluctuations and the allocation of time", Journal of Labor Economics 15:S223–S250.
Hall, R.E., and E.P. Lazear (1984), "The excess sensitivity of layoffs and quits to demand", Journal of Labor Economics 2:233–257.
Hall, R.E., and D.M. Lilien (1979), "Efficient wage bargains under uncertain supply and demand", American Economic Review 69:868–879.
Hart, O. (1983), "Optimal labour contracts under asymmetric information: an introduction", Review of Economic Studies 50:3–35.
Hashimoto, M., and B. Yu (1980), "Specific capital, employment contracts, and wage rigidity", Bell Journal of Economics 11:536–549.
Heckman, J.J., and B. Singer (1985), Longitudinal Analysis of Labor Market Data (Cambridge University Press, New York).
Kahn, C., and G. Huberman (1988), "Two-sided uncertainty and 'up-or-out' contracts", Journal of Labor Economics 6:423–444.
Killingsworth, M.R., and J.J. Heckman (1986), "Female labor supply: a survey", in: O. Ashenfelter and R. Layard, eds., Handbook of Labor Economics, vol. 1 (North-Holland, Amsterdam) ch. 2, 103–204.
King, R.G., and S.T. Rebelo (1999), "Resuscitating real business cycles", ch. 14, this Handbook.
Kydland, F.E., and E.C. Prescott (1982), "Time to build and aggregate fluctuations", Econometrica 5:1345–1370.
Lilien, D.M. (1982), "Sectoral shifts and cyclical unemployment", Journal of Political Economy 90: 777–793.
Lucas, R.E. (1972), "Expectations and the neutrality of money", Journal of Economic Theory 4:103–124.
MacLeod, W.B., J.M. Malcomson and P. Gomme (1994), "Labor turnover and the natural rate of unemployment: efficiency wage versus frictional unemployment", Journal of Labor Economics 12: 276–315.
MaCurdy, T.E. (1981), "An empirical model of labor supply in a life cycle setting", Journal of Political Economy 89:1059–1085.
McLaughlin, K.J. (1991), "A theory of quits and layoffs with efficient turnover", Journal of Political Economy 99:1–29.
Merz, M. (1995), "Search in the labor market and the real business cycle", Journal of Monetary Economics 36:269–300.
Mirrlees, J.A. (1971), "An exploration in the theory of optimum income taxation", Review of Economic Studies 38:175–208.
Mortensen, D.T. (1982), "Property rights and efficiency in mating, racing, and related games", American Economic Review 72:968–979.

Mortensen, D.T. (1986), "Job search and labor market analysis", in: O. Ashenfelter and R. Layard, eds., Handbook of Labor Economics (North-Holland, Amsterdam) 849–919.
Mortensen, D.T., and C.A. Pissarides (1994), "Job creation and destruction in the theory of unemployment", Review of Economic Studies 61:397–415.
Mulligan, C.B. (1995), "The intertemporal substitution of work – what does the evidence say", University of Chicago Population Research Center Discussion Paper 95-3.
Pencavel, J. (1986), "Labor supply of men: a survey", in: O. Ashenfelter and R. Layard, eds., Handbook of Labor Economics (North-Holland, Amsterdam) 3–102.
Phelps, E.S. (1994), Structural Slumps: The Modern Equilibrium Theory of Unemployment, Interest, and Assets (Harvard University Press, Cambridge, MA).
Picard, P. (1993), Wages and Unemployment: A Study in Non-Walrasian Macroeconomics (Cambridge University Press, Cambridge, MA).
Pissarides, C.A. (1990), Equilibrium Unemployment Theory (Blackwell, Oxford).
Ramey, G., and J. Watson (1997), "Contractual fragility, job destruction, and business cycles", Quarterly Journal of Economics 112:873–911.
Rogerson, R. (1988), "Indivisible labor, lotteries, and equilibrium", Journal of Monetary Economics 21:3–16.
Romer, D. (1996), Advanced Macroeconomics (McGraw-Hill, New York).
Rotemberg, J.J., and M. Woodford (1996), "Real-business-cycle models and the forecastable movements in output, hours, and consumption", American Economic Review 86:71–89.
Ruhm, C. (1991), "Are workers permanently scarred by job displacements?", American Economic Review 81:319–324.
Saint-Paul, G. (1996), "Efficiency wages as a persistence mechanism", in: H.D. Dixon and N. Rankin, eds., The New Macroeconomics: Imperfect Markets and Policy Effectiveness (Cambridge University Press, Cambridge) 186–205.
Sargent, T.J. (1979), Macroeconomic Theory (Academic Press, New York).
Segal, I.B., and M.D. Whinston (1996), "Naked exclusive contracts and buyer coordination", Discussion Paper 1780 (Harvard Institute of Economic Research).
Segal, I.B., and M.D. Whinston (1997), Exclusive Dealing and Specific Investments (Harvard University and University of California, Berkeley).
Shapiro, C., and J.E. Stiglitz (1984), "Equilibrium unemployment as a worker discipline device", American Economic Review 74:433–444.
Solow, R.M. (1957), "Technical change and the aggregate production function", Review of Economics and Statistics 39:312–320.
Woodford, M. (1994a), "Notes on dynamic efficiency wage models". Unpublished.
Woodford, M. (1994b), "Structural slumps", Journal of Economic Literature 32:1784–1815.
Wright, R. (1986), "Job search and cyclical unemployment", Journal of Political Economy 94:38–55.