Chapter 10
UNEMPLOYMENT

10.1 Introduction: Theories of Unemployment

In almost any economy at almost any time, many individuals appear to be unemployed. That is, there are many people who are not working but who say they want to work in jobs like those held by individuals similar to them, at the wages those individuals are earning.

The possibility of unemployment is a central subject of macroeconomics. There are two basic issues. The first concerns the determinants of average unemployment over extended periods. The central questions here are whether this unemployment represents a genuine failure of markets to clear, and if so, what its causes and consequences are. There is a wide range of possible views. At one extreme is the position that unemployment is largely illusory, or the working out of unimportant frictions in the process of matching up workers and jobs. At the other extreme is the view that unemployment is the result of non-Walrasian features of the economy and that it largely represents a waste of resources.

The second issue concerns the cyclical behavior of the labor market. As described in Section 6.3, the real wage appears to be only moderately procyclical. This is consistent with the view that the labor market is Walrasian only if labor supply is quite elastic or if shifts in labor supply play an important role in employment fluctuations. But as we saw in Section 5.10, there is little support for the hypothesis of highly elastic labor supply. And it seems unlikely that shifts in labor supply are central to fluctuations. The remaining possibility is that the labor market is not Walrasian, and that its non-Walrasian features are central to its cyclical behavior. That possibility is the focus of this chapter.

The issue of why shifts in labor demand appear to lead to large movements in employment and only small movements in the real wage is important to all theories of fluctuations. For example, we saw in Chapter 6 that if the real wage is highly procyclical in response to demand shocks, it is essentially impossible for the small barriers to nominal adjustment to

generate substantial nominal rigidity. In the face of a decline in aggregate demand, for example, if prices remain fixed the real wage must fall sharply; as a result, each firm has a huge incentive to cut its price and hire labor to produce additional output. If, however, there is some non-Walrasian feature of the labor market that causes the cost of labor to respond little to the overall level of economic activity, then there is some hope for theories of small frictions in nominal adjustment.

This chapter considers various ways in which the labor market may depart from a competitive, textbook market. We investigate both whether these departures can lead to substantial unemployment and whether they can have large effects on the cyclical behavior of employment and the real wage.

If there is unemployment in a Walrasian labor market, unemployed workers immediately bid the wage down until supply and demand are in balance. Theories of unemployment can therefore be classified according to their view of why this mechanism fails to operate. Concretely, consider an unemployed worker who offers to work for a firm for slightly less than the firm is currently paying, and who is otherwise identical to the firm's current workers. There are at least four possible responses the firm can make to this offer.

First, the firm can say that it does not want to reduce wages. Theories in which there is a cost as well as a benefit to the firm of paying lower wages are known as *efficiency-wage* theories. (The name comes from the idea that higher wages may raise the productivity, or efficiency, of labor.) These theories are the subject of Sections 10.2 through 10.4. Section 10.2 first discusses the possible ways that paying lower wages can harm a firm; it then analyzes a simple model where wages affect productivity but where the reason for that link is not explicitly specified. Section 10.3 considers an important generalization of that model. Finally, Section 10.4 presents a model formalizing one particular view of why paying higher wages can be beneficial. The central idea is that if firms cannot monitor their workers' effort perfectly, they may pay more than market-clearing wages to induce workers not to shirk.

The second possible response the firm can make is that it wishes to cut wages, but that an explicit or implicit agreement with its workers prevents it from doing so.[1] Theories in which bargaining and contracts affect the macroeconomics of the labor market are known as *contracting models*. These models are considered in Section 10.5.

The third way the firm can respond to the unemployed worker's offer is to say that it does not accept the premise that the unemployed worker is identical to the firm's current employees. That is, heterogeneity among workers and jobs may be an essential feature of the labor market. In this

[1] The firm can also be prevented from cutting wages by minimum-wage laws. In most settings, this is relevant only to low-skill workers; thus it does not appear to be central to the macroeconomics of unemployment.

view, to think of the market for labor as a single market, or even as a large number of interconnected markets, is to commit a fundamental error. Instead, according to this view, each worker and each job should be thought of as distinct; as a result, the process of matching up workers and jobs occurs not through markets but through a complex process of search. Models of this type are known as *search and matching models*. They are discussed in Sections 10.6 and 10.7.

Finally, the firm can accept the worker's offer. That is, it is possible that the market for labor is approximately Walrasian. In this view, measured unemployment consists largely of people who are moving between jobs, or who would like to work at wages higher than those they can in fact obtain. Since the focus of this chapter is on unemployment, we will not develop this idea here. Nonetheless, it is important to keep in mind that this is one view of the labor market.

10.2 A Generic Efficiency-Wage Model

Potential Reasons for Efficiency Wages

The key assumption of efficiency-wage models is that there is a benefit as well as a cost to a firm of paying a higher wage. There are many reasons that this could be the case. Here we describe four of the most important.

First, and most simply, a higher wage can increase workers' food consumption, and thereby cause them to be better nourished and more productive. Obviously this possibility is not important in developed economies. Nonetheless, it provides a concrete example of an advantage of paying a higher wage. For that reason, it is often a useful reference point.

Second, a higher wage can increase workers' effort in situations where the firm cannot monitor them perfectly. In a Walrasian labor market, workers are indifferent about losing their jobs, since identical jobs are immediately available. Thus if the only way that firms can punish workers who exert low effort is by firing them, workers in such a labor market have no incentive to exert effort. But if a firm pays more than the market-clearing wage, its jobs are valuable. Thus its workers may choose to exert effort even if there is some chance they will not be caught if they shirk. This idea is developed in Section 10.4.

Third, paying a higher wage can improve workers' ability along dimensions the firm cannot observe. Specifically, if higher-ability workers have higher reservation wages, offering a higher wage raises the average quality of the applicant pool, and thus raises the average ability of the workers the firm hires (Weiss, 1980).[2]

[2] When ability is observable, the firm can pay higher wages to more able workers. Thus observable ability differences do not lead to any departures from the Walrasian case.

Finally, a high wage can build loyalty among workers and hence induce high effort; conversely, a low wage can cause anger and desire for revenge, and thereby lead to shirking or sabotage. Akerlof and Yellen (1990) present extensive evidence that workers' effort is affected by such forces as anger, jealousy, and gratitude. For example, they describe studies showing that workers who believe they are underpaid sometimes perform their work in ways that are harder for them in order to reduce their employers' profits.[3]

Other Compensation Schemes

This discussion implicitly assumes that a firm's financial arrangements with its workers take the form of some wage per unit of time. An important question is whether there are more complicated ways for the firm to compensate its workers that allow it to obtain the benefits of a higher wage less expensively. The nutritional advantages of a higher wage, for example, can be obtained by compensating workers partly in kind (such as by feeding them at work). To give another example, firms can give workers an incentive to exert effort by requiring them to post a bond that they lose if they are caught shirking.

If there are cheaper ways for firms to obtain the benefits of a higher wage, then these benefits lead not to a higher wage but just to complicated compensation policies. Whether the benefits can be obtained in such ways depends on the specific reason that a higher wage is advantageous. We will therefore not attempt a general treatment. The end of Section 10.4 discusses this issue in the context of efficiency-wage theories based on imperfect monitoring of workers' effort. In this section and the next, however, we simply assume that compensation takes the form of a conventional wage, and investigate the effects of efficiency wages under this assumption.

Assumptions

We now turn to a model of efficiency wages. There is a large number, N, of identical competitive firms.[4] The representative firm seeks to maximize its profits, which are given by

$$\pi = Y - wL, \tag{10.1}$$

[3] See Problem 10.5 for a formalization of this idea. Three other potential advantages of a higher wage are that it can reduce turnover (and hence recruitment and training costs, if they are borne by the firm); that it can lower the likelihood that the workers will unionize; and that it can raise the utility of managers who have some ability to pursue objectives other than maximizing profits.

[4] We can think of the number of firms as being determined by the amount of capital in the economy, which is fixed in the short run.

where Y is the firm's output, w is the wage that it pays, and L is the amount of labor it hires.

A firm's output depends on the number of workers it employs and on their effort. For simplicity, we neglect other inputs and assume that labor and effort enter the production function multiplicatively. Thus the representative firm's output is

$$Y = F(eL), \qquad F'(\bullet) > 0, \qquad F''(\bullet) < 0, \tag{10.2}$$

where e denotes workers' effort. The crucial assumption of efficiency-wage models is that effort depends positively on the wage the firm pays. In this section we consider the simple case (due to Solow, 1979) where the wage is the only determinant of effort. Thus,

$$e = e(w), \qquad e'(\bullet) > 0. \tag{10.3}$$

Finally, there are $\overline{L}$ identical workers, each of whom supplies 1 unit of labor inelastically.

Analyzing the Model

The problem facing the representative firm is

$$\max_{L,w} F(e(w)L) - wL. \tag{10.4}$$

If there are unemployed workers, the firm can choose the wage freely. If unemployment is zero, on the other hand, the firm must pay at least the wage paid by other firms.

When the firm is unconstrained, the first-order conditions for L and w are[5]

$$F'(e(w)L)e(w) - w = 0, \tag{10.5}$$

$$F'(e(w)L)Le'(w) - L = 0. \tag{10.6}$$

We can rewrite (10.5) as

$$F'(e(w)L) = \frac{w}{e(w)}. \tag{10.7}$$

Substituting (10.7) into (10.6) and dividing by L yields

$$\frac{we'(w)}{e(w)} = 1. \tag{10.8}$$

Equation (10.8) states that at the optimum, the elasticity of effort with respect to the wage is 1. To understand this condition, note that output is a function of the quantity of effective labor, eL. The firm therefore wants to hire effective labor as cheaply as possible. When the firm hires a worker, it

[5] We assume that the second-order conditions are satisfied.

obtains $e(w)$ units of effective labor at a cost of w; thus the cost per unit of effective labor is $w/e(w)$. When the elasticity of e with respect to w is 1, a marginal change in w has no effect on this ratio; thus this is the first-order condition for the problem of choosing w to minimize the cost of effective labor. The wage satisfying (10.8) is known as the *efficiency wage*.

Figure 10.1 depicts the choice of w graphically in (w,e) space. The rays coming out from the origin are lines where the ratio of e to w is constant; the ratio is larger on the higher rays. Thus the firm wants to choose w to attain as high a ray as possible. This occurs where the $e(w)$ function is just tangent to one of the rays—that is, where the elasticity of e with respect to w is 1. Panel (a) shows a case where effort is sufficiently responsive to the wage that over some range the firm prefers a higher wage. Panel (b) shows a case where the firm always prefers a lower wage.

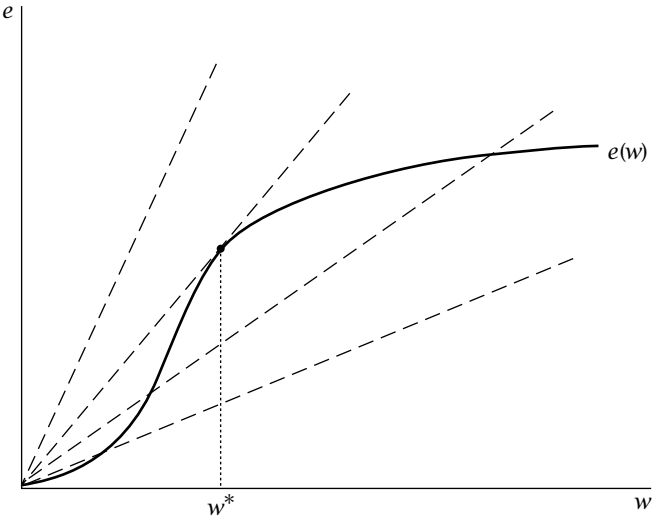
Finally, equation (10.7) states that the firm hires workers until the marginal product of effective labor equals its cost. This is analogous to the condition in a standard labor-demand problem that the firm hires labor up to the point where the marginal product equals the wage.

Equations (10.7) and (10.8) describe the behavior of a single firm. Describing the economy-wide equilibrium is straightforward. Let w^* and L^* denote the values of w and L that satisfy (10.7) and (10.8). Since firms are identical, each firm chooses these same values of w and L. Total labor demand is therefore NL^*. If labor supply, $\overline{L}$, exceeds this amount, firms are unconstrained in their choice of w. In this case the wage is w^*, employment is NL^*, and there is unemployment of amount $\overline{L} - NL^*$. If NL^* exceeds $\overline{L}$, on the other hand, firms are constrained. In this case, the wage is bid up to the point where demand and supply are in balance, and there is no unemployment.

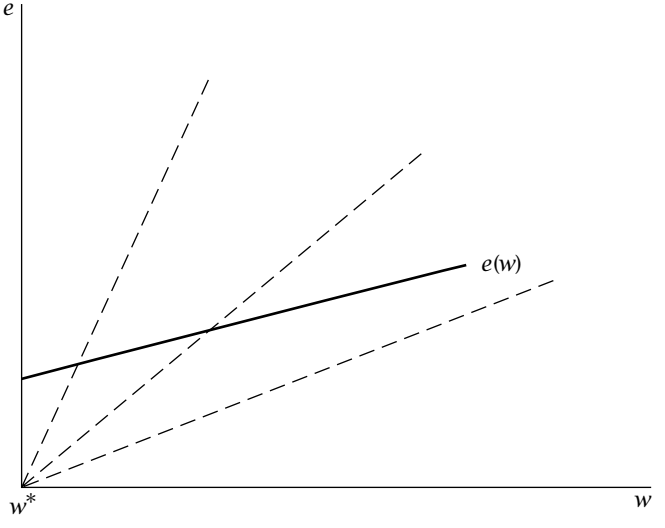
Implications

This model shows how efficiency wages can give rise to unemployment. In addition, the model implies that the real wage is unresponsive to demand shifts. Suppose the demand for labor increases. Since the efficiency wage, w^*, is determined entirely by the properties of the effort function, $e(\bullet)$, there is no reason for firms to adjust their wages. Thus the model provides a candidate explanation of why shifts in labor demand lead to large movements in employment and small changes in the real wage. In addition, the fact that the real wage and effort do not change implies that the cost of a unit of effective labor does not change. As a result, in a model with price-setting firms, the incentive to adjust prices is small.

Unfortunately, these results are less promising than they appear. The difficulty is that they apply not just to the short run but to the long run: the model implies that as economic growth shifts the demand for labor outward, the real wage remains unchanged and unemployment trends downward. Eventually, unemployment reaches zero, at which point further increases in



(a)



(b)

FIGURE 10.1 The determination of the efficiency wage

demand lead to increases in the real wage. In practice, however, we observe no clear trend in unemployment over extended periods. In other words, the basic fact about the labor market that we need to understand is not just that shifts in labor demand appear to have little impact on the real wage and fall mainly on employment in the short run; it is also that they fall almost entirely on the real wage in the long run. Our model does not explain this pattern.

10.3 A More General Version

Introduction

With many of the potential sources of efficiency wages, the wage is unlikely to be the only determinant of effort. Suppose, for example, that the wage affects effort because firms cannot monitor workers perfectly and workers are concerned about the possibility of losing their jobs if the firm catches them shirking. In such a situation, the cost to a worker of being fired depends not just on the wage the job pays, but also on how easy it is to obtain other jobs and on the wages those jobs pay. Thus workers are likely to exert more effort at a given wage when unemployment is higher, and to exert less effort when the wage paid by other firms is higher. Similar arguments apply to situations where the wage affects effort because of unobserved ability or feelings of gratitude or anger.

Thus a natural generalization of the effort function, (10.3), is

$$e = e(w, w_a, u), \qquad e_1(\bullet) > 0, \qquad e_2(\bullet) < 0, \qquad e_3(\bullet) > 0, \tag{10.9}$$

where w_a is the wage paid by other firms and u is the unemployment rate, and where subscripts denote partial derivatives.

Each firm is small relative to the economy, and therefore takes w_a and u as given. The representative firm's problem is the same as before, except that w_a and u now affect the effort function. The first-order conditions can therefore be rearranged to obtain

$$F'(e(w, w_a, u)L) = \frac{w}{e(w, w_a, u)}, \tag{10.10}$$

$$\frac{we_1(w, w_a, u)}{e(w, w_a, u)} = 1. \tag{10.11}$$

These conditions are analogous to (10.7) and (10.8) in the simpler version of the model.

Assume that the $e(\bullet)$ function is sufficiently well behaved that there is a unique optimal w for a given w_a and u. Given this assumption, equilibrium requires $w = w_a$; if not, each firm wants to pay a wage different from the

prevailing wage. Let w^* and L^* denote the values of w and L satisfying (10.10)–(10.11) with $w = w_a$. As before, if NL^* is less than $\overline{L}$, the equilibrium wage is w^* and $\overline{L} - NL^*$ workers are unemployed. And if NL^* exceeds $\overline{L}$, the wage is bid up and the labor market clears.

This extended version of the model has promise for accounting for both the absence of any trend in unemployment over the long run and the fact that shifts in labor demand appear to have large effects on unemployment in the short run. This is most easily seen by means of an example.

Example

Following Summers (1988), suppose that effort is given by

$$e = \begin{cases} \left(\dfrac{w - x}{x}\right)^{\beta} & \text{if } w > x \\ 0 & \text{otherwise,} \end{cases} \tag{10.12}$$

$$x = (1 - bu)w_a, \tag{10.13}$$

where $0 < \beta < 1$ and $b > 0$. x is a measure of labor-market conditions. If b equals 1, x is the wage paid at other firms multiplied by the fraction of workers who are employed. If b is less than 1, workers put less weight on unemployment; this could occur if there are unemployment benefits or if workers value leisure. If b is greater than 1, workers put more weight on unemployment; this might occur because workers who lose their jobs face unusually high chances of continued unemployment, or because of risk aversion. Finally, equation (10.12) states that for $w > x$, effort increases less than proportionately with $w - x$.

Differentiation of (10.12) shows that for this functional form, the condition that the elasticity of effort with respect to the wage equals 1 (equation [10.11]) is

$$\beta \frac{w}{[(w - x)/x]^{\beta}} \left(\frac{w - x}{x}\right)^{\beta - 1} \frac{1}{x} = 1. \tag{10.14}$$

Straightforward algebra can be used to simplify (10.14) to

$$\begin{aligned} w &= \frac{x}{1 - \beta} \\ &= \frac{1 - bu}{1 - \beta} w_a. \end{aligned} \tag{10.15}$$

For small values of β, $1/(1 - \beta) \simeq 1 + \beta$. Thus (10.15) implies that when β is small, the firm offers a premium of approximately fraction β over the index of labor-market opportunities, x.

Equilibrium requires that the representative firm wants to pay the prevailing wage, or that $w = w_a$. Imposing this condition in (10.15) yields

$$(1 - \beta)w_a = (1 - bu)w_a. \tag{10.16}$$

For this condition to be satisfied, the unemployment rate must be given by

$$\begin{aligned} u &= \frac{\beta}{b} \\ &\equiv u_{EQ}. \end{aligned} \tag{10.17}$$

As equation (10.15) shows, each firm wants to pay more than the prevailing wage if unemployment is less than u_{EQ}, and wants to pay less if unemployment is more than u_{EQ}. Thus equilibrium requires that $u = u_{EQ}$.

Implications

This analysis has three important implications. First, (10.17) implies that equilibrium unemployment depends only on the parameters of the effort function; the production function is irrelevant. Thus an upward trend in the production function does not produce a trend in unemployment.

Second, relatively modest values of β—the elasticity of effort with respect to the premium firms pay over the index of labor-market conditions—can lead to nonnegligible unemployment. For example, either $\beta = 0.06$ and $b = 1$ or $\beta = 0.03$ and $b = 0.5$ imply that equilibrium unemployment is 6 percent. This result is not as strong as it may appear, however: while these parameter values imply a low elasticity of effort with respect to $(w - x)/x$, they also imply that workers exert no effort at all until the wage is quite high. For example, if b is 0.5 and unemployment is at its equilibrium level of 6 percent, effort is zero until a firm's wage reaches 97 percent of the prevailing wage. In that sense, efficiency-wage forces are quite strong for these parameter values.

Third, firms' incentive to adjust wages or prices (or both) in response to changes in aggregate unemployment is likely to be small for reasonable cases. Suppose we embed this model of wages and effort in a model of price-setting firms along the lines of Chapter 6. Consider a situation where the economy is initially in equilibrium, so that $u = u_{EQ}$ and marginal revenue and marginal cost are equal for the representative firm. Now suppose that the money supply falls and firms do not change their nominal wages or prices; as a result, unemployment rises above u_{EQ}. We know from Chapter 6 that small barriers to wage and price adjustment can cause this to be an equilibrium only if the representative firm's incentive to adjust is small.

For concreteness, consider the incentive to adjust wages. Equation (10.15), $w = (1 - bu)w_a/(1 - \beta)$, shows that the cost-minimizing wage is decreasing in the unemployment rate. Thus the firm can reduce its costs, and hence raise its profits, by cutting its wage. The key issue is the size of the gain. Equation (10.12) for effort implies that if the firm leaves its wage equal to

the prevailing wage, w_a, its cost per unit of effective labor, w/e, is

$$\begin{aligned} C_{\text{FIXED}} &= \frac{w_a}{e(w_a, w_a, u)} \\ &= \frac{w_a}{\left(\dfrac{w_a - x}{x}\right)^{\beta}} \\ &= \frac{w_a}{\left[\dfrac{w_a - (1 - bu)w_a}{(1 - bu)w_a}\right]^{\beta}} \\ &= \left(\frac{1 - bu}{bu}\right)^{\beta} w_a. \end{aligned} \tag{10.18}$$

If the firm changes its wage, on the other hand, it sets it according to (10.15), and thus chooses $w = x/(1 - \beta)$. In this case, the firm's cost per unit of effective labor is

$$\begin{aligned} C_{\text{ADJ}} &= \frac{w}{\left(\dfrac{w - x}{x}\right)^{\beta}} \\ &= \frac{x/(1 - \beta)}{\left\{\dfrac{[x/(1 - \beta)] - x}{x}\right\}^{\beta}} \\ &= \frac{x/(1 - \beta)}{\left[\beta/(1 - \beta)\right]^{\beta}} \\ &= \frac{1}{\beta^{\beta}} \frac{1}{(1 - \beta)^{1-\beta}} (1 - bu)w_a. \end{aligned} \tag{10.19}$$

Suppose that $\beta = 0.06$ and $b = 1$, so that $u_{\text{EQ}} = 6\%$. Suppose, however, that unemployment rises to 9 percent and that other firms do not change their wages. Equations (10.18) and (10.19) imply that this rise lowers C_{FIXED} by 2.6 percent and C_{ADJ} by 3.2 percent. Thus the firm can save only 0.6 percent of costs by cutting its wages. For $\beta = 0.03$ and $b = 0.5$, the declines in C_{FIXED} and C_{ADJ} are 1.3 percent and 1.5 percent; thus in this case the incentive to cut wages is even smaller.[6]

[6] One can also show that if firms do not change their wages, for reasonable cases their incentive to adjust their prices is also small. If wages are completely flexible, however, the incentive to adjust prices is not small. With u greater than u_{EQ}, each firm wants to pay less than other firms are paying (see [10.15]). Thus if wages are completely flexible, they must fall to zero—or, if workers have a positive reservation wage, to the reservation wage. As a result, firms' labor costs are extremely low, and so their incentive to cut prices and increase output is high. Thus in the absence of any barriers to changing wages, small costs to changing prices are not enough to prevent price adjustment in this model.

In a competitive labor market, in contrast, the equilibrium wage falls by the percentage fall in employment divided by the elasticity of labor supply. For a 3 percent fall in employment and a labor supply elasticity of 0.2, for example, the equilibrium wage falls by 15 percent. And without endogenous effort, a 15 percent fall in wages translates directly into a 15 percent fall in costs. Firms therefore have an overwhelming incentive to cut wages and prices in this case.[7]

Thus efficiency wages have a potentially large impact on the incentive to adjust wages in the face of fluctuations in aggregate output. As a result, they have the potential to explain why shifts in labor demand mainly affect employment in the short run. Intuitively, in a competitive market firms are initially at a corner solution with respect to wages: firms pay the lowest possible wage at which they can hire workers. Thus wage reductions, if possible, are unambiguously beneficial. With efficiency wages, in contrast, firms are initially at an interior optimum where the marginal benefits and costs of wage cuts are equal.

10.4 The Shapiro–Stiglitz Model

One source of efficiency wages that has received a great deal of attention is the possibility that firms' limited monitoring abilities force them to provide their workers with an incentive to exert effort. This section presents a specific model, due to Shapiro and Stiglitz (1984), of this possibility.

Presenting a formal model of imperfect monitoring serves three purposes. First, it allows us to investigate whether this idea holds up under scrutiny. Second, it permits us to analyze additional questions. For example, only with a formal model can we ask whether government policies can improve welfare. Third, the mathematical tools the model employs are useful in other settings.

Assumptions

The economy consists of a large number of workers, $\overline{L}$, and a large number of firms, N. Workers maximize their expected discounted utilities, and firms maximize their expected discounted profits. The model is set in continuous time. For simplicity, the analysis focuses on steady states.

[7] In fact, in a competitive labor market, an individual firm's incentive to reduce wages if other firms do not is even larger than the fall in the equilibrium wage. If other firms do not cut wages, some workers are unemployed. Thus the firm can hire workers at an arbitrarily small wage (or at workers' reservation wage).

Consider workers first. The representative worker's lifetime utility is

$$U = \int_{t=0}^{\infty} e^{-\rho t} u(t) dt, \qquad \rho > 0. \tag{10.20}$$

$u(t)$ is instantaneous utility at time t, and ρ is the discount rate. Instantaneous utility is

$$u(t) = \begin{cases} w(t) - e(t) & \text{if employed} \\ 0 & \text{if unemployed.} \end{cases} \tag{10.21}$$

w is the wage and e is the worker's effort. There are only two possible effort levels, $e = 0$ and $e = \bar{e}$. Thus at any moment a worker must be in one of three states: employed and exerting effort (denoted E), employed and not exerting effort (denoted S, for shirking), or unemployed (denoted U).

A key ingredient of the model is its assumptions concerning workers' transitions among the three states. First, there is an exogenous rate at which jobs end. Specifically, if a worker begins working in a job at some time, t_0 (and if the worker exerts effort), the probability that the worker is still employed in the job at some later time, t, is

$$P(t) = e^{-b(t-t_0)}, \qquad b > 0. \tag{10.22}$$

(10.22) implies that $P(t+\tau)/P(t)$ equals $e^{-b\tau}$, and thus that it is independent of t: if a worker is employed at some time, the probability that he or she is still employed time τ later is $e^{-b\tau}$ regardless of how long the worker has already been employed. This assumption that job breakups follow a Poisson process simplifies the analysis greatly, because it implies that there is no need to keep track of how long workers have been in their jobs.

An equivalent way to describe the process of job breakup is to say that it occurs with probability b per unit time, or to say that the *hazard rate* for job breakup is b. That is, the probability that an employed worker's job ends in the next dt units of time approaches bdt as dt approaches zero. To see that our assumptions imply this, note that (10.22) implies $P'(t) = -bP(t)$.

The second assumption concerning workers' transitions between states is that firms' detection of workers who are shirking is also a Poisson process. Specifically, detection occurs with probability q per unit time. q is exogenous, and detection is independent of job breakups. Workers who are caught shirking are fired. Thus if a worker is employed but shirking, the probability that he or she is still employed time τ later is $e^{-q\tau}$ (the probability that the worker has not been caught and fired) times $e^{-b\tau}$ (the probability that the job has not ended exogenously).

Third, unemployed workers find employment at rate a per unit time. Each worker takes a as given. In the economy as a whole, however, a is determined endogenously. When firms want to hire workers, they choose workers at random out of the pool of unemployed workers. Thus a is determined by the rate at which firms are hiring (which is determined by the number of employed workers and the rate at which jobs end) and the number of

unemployed workers. Because workers are identical, the probability of finding a job does not depend on how workers become unemployed or on how long they are unemployed.

Firms' behavior is straightforward. A firm's profits at t are

$$\pi(t) = F(\bar{e}L(t)) - w(t)[L(t) + S(t)], \qquad F'(\bullet) > 0, \qquad F''(\bullet) < 0, \qquad (10.23)$$

where L is the number of employees who are exerting effort and S is the number who are shirking. The problem facing the firm is to set w sufficiently high that its workers do not shirk, and to choose L. Because the firm's decisions at any date affect profits only at that date, there is no need to analyze the present value of profits: the firm chooses w and L at each moment to maximize the instantaneous flow of profits.

The final assumption of the model is $\bar{e}F'(\bar{e}\overline{L}/N) > \bar{e}$, or $F'(\bar{e}\overline{L}/N) > 1$. This condition states that if each firm hires $1/N$ of the labor force, the marginal product of labor exceeds the cost of exerting effort. Thus in the absence of imperfect monitoring, there is full employment.

The Values of *E*, *U*, and *S*

Let V_i denote the "value" of being in state i (for $i = E$, S, and U). That is, V_i is the expected value of discounted lifetime utility from the present moment forward of a worker who is in state i. Because transitions among states are Poisson processes, the V_i's do not depend on how long the worker has been in the current state or on the worker's prior history. And because we are focusing on steady states, the V_i's are constant over time.

To find V_E, V_S, and V_U, it is not necessary to analyze the various paths the worker may follow over the infinite future. Instead we can use *dynamic programming*. The central idea of dynamic programming is to look at only a brief interval of time and use the V_i's themselves to summarize what occurs after the end of the interval.[8] Consider first a worker who is employed and exerting effort at time 0. Suppose temporarily that time is divided into intervals of length Δt, and that a worker who loses his or her job during one interval cannot begin to look for a new job until the beginning of the next interval. Let $V_E(\Delta t)$ and $V_U(\Delta t)$ denote the values of employment and unemployment as of the beginning of an interval under this assumption. In a moment we will let Δt approach zero. When we do this, the constraint that a worker who loses his or her job during an interval cannot find a new job during the remainder of that interval becomes irrelevant. Thus $V_E(\Delta t)$ will approach V_E.

[8] If time is discrete rather than continuous, we look one period ahead. See Ljungqvist and Sargent (2004) for an introduction to dynamic programming.

If a worker is employed in a job paying a wage of w, $V_E(\Delta t)$ is given by

$$V_E(\Delta t) = \int_{t=0}^{\Delta t} e^{-bt}e^{-\rho t}(w-\bar{e})\,dt + e^{-\rho\Delta t}[e^{-b\Delta t}V_E(\Delta t) + (1-e^{-b\Delta t})V_U(\Delta t)]. \tag{10.24}$$

The first term of (10.24) reflects utility during the interval $(0,\Delta t)$. The probability that the worker is still employed at time t is e^{-bt}. If the worker is employed, flow utility is $w-\bar{e}$. Discounting this back to time 0 yields an expected contribution to lifetime utility of $e^{-(\rho+b)t}(w-\bar{e})$.[9]

The second term of (10.24) reflects utility after Δt. At time Δt, the worker is employed with probability $e^{-b\Delta t}$ and unemployed with probability $1-e^{-b\Delta t}$. Combining these probabilities with the V's and discounting yields the second term.

If we compute the integral in (10.24), we can rewrite the equation as

$$V_E(\Delta t) = \frac{1}{\rho+b}\left(1-e^{-(\rho+b)\Delta t}\right)(w-\bar{e}) + e^{-\rho\Delta t}[e^{-b\Delta t}V_E(\Delta t) + (1-e^{-b\Delta t})V_U(\Delta t)]. \tag{10.25}$$

Solving this expression for $V_E(\Delta t)$ gives

$$V_E(\Delta t) = \frac{1}{\rho+b}(w-\bar{e}) + \frac{1}{1-e^{-(\rho+b)\Delta t}}e^{-\rho\Delta t}(1-e^{-b\Delta t})V_U(\Delta t). \tag{10.26}$$

As described above, V_E equals the limit of $V_E(\Delta t)$ as Δt approaches zero. (Similarly, V_U equals the limit of $V_U(\Delta t)$ as t approaches zero.) To find this limit, we apply l'Hôpital's rule to (10.26). This yields

$$V_E = \frac{1}{\rho+b}[(w-\bar{e}) + bV_U]. \tag{10.27}$$

Equation (10.27) can also be derived intuitively. Think of an asset that pays dividends at rate $w-\bar{e}$ per unit time when the worker is employed and no dividends when the worker is unemployed. In addition, assume that the asset is being priced by risk-neutral investors with required rate of return ρ. Since the expected present value of lifetime dividends of this asset is the same as the worker's expected present value of lifetime utility, the asset's price must be V_E when the worker is employed and V_U when the worker is unemployed. For the asset to be held, it must provide an expected rate of return of ρ. That is, its dividends per unit time, plus any expected capital gains or losses per unit time, must equal ρV_E. When the worker is employed, dividends per unit time are $w-\bar{e}$, and there is a probability b per unit time

[9] Because of the steady-state assumption, if it is optimal for the worker to exert effort initially, it continues to be optimal. Thus we do not have to allow for the possibility of the worker beginning to shirk.

of a capital loss of $V_E - V_U$. Thus,

$$\rho V_E = (w - \bar{e}) - b(V_E - V_U). \tag{10.28}$$

Rearranging this expression yields (10.27).

If the worker is shirking, the "dividend" is w per unit time, and the expected capital loss is $(b + q)(V_S - V_U)$ per unit time. Thus reasoning parallel to that used to derive (10.28) implies

$$\rho V_S = w - (b + q)(V_S - V_U). \tag{10.29}$$

Finally, if the worker is unemployed, the dividend is zero and the expected capital gain (assuming that firms pay sufficiently high wages that employed workers exert effort) is $a(V_E - V_U)$ per unit time.[10] Thus,

$$\rho V_U = a(V_E - V_U). \tag{10.30}$$

The No-Shirking Condition

The firm must pay enough that $V_E \geq V_S$; otherwise its workers exert no effort and produce nothing. At the same time, since effort cannot exceed $\bar{e}$, there is no need to pay any excess over the minimum needed to induce effort. Thus the firm chooses w so that V_E just equals V_S:[11]

$$V_E = V_S. \tag{10.31}$$

This result tells us that the left-hand sides of (10.28) and (10.29) must be equal. Thus

$$(w - \bar{e}) - b(V_E - V_U) = w - (b + q)(V_E - V_U), \tag{10.32}$$

or

$$V_E - V_U = \frac{\bar{e}}{q}. \tag{10.33}$$

Equation (10.33) implies that firms set wages high enough that workers strictly prefer employment to unemployment. Thus workers obtain rents. The size of the premium is increasing in the cost of exerting effort, $\bar{e}$, and decreasing in firms' efficacy in detecting shirkers, q.

The next step is to find what the wage must be for the rent to employment to equal $\bar{e}/q$. Equations (10.28) and (10.30) imply

$$\rho(V_E - V_U) = (w - \bar{e}) - (a + b)(V_E - V_U). \tag{10.34}$$

[10] Equations (10.29) and (10.30) can also be derived by defining $V_U(\Delta t)$ and $V_S(\Delta t)$ and proceeding along the lines used to derive (10.27).

[11] Since all firms are the same, they choose the same wage.

It follows that for $V_E - V_U$ to equal $\bar{e}/q$, the wage must satisfy

$$w = \bar{e} + (a + b + \rho)\frac{\bar{e}}{q}. \tag{10.35}$$

Thus the wage needed to induce effort is increasing in the cost of effort ($\bar{e}$), the ease of finding jobs (a), the rate of job breakup (b), and the discount rate (ρ), and decreasing in the probability that shirkers are detected (q).

It turns out to be more convenient to express the wage needed to prevent shirking in terms of employment per firm, L, rather than the rate at which the unemployed find jobs, a. To substitute for a, we use the fact that, since the economy is in steady state, movements into and out of unemployment balance. The number of workers becoming unemployed per unit time is N (the number of firms) times L (the number of workers per firm) times b (the rate of job breakup).[12] The number of unemployed workers finding jobs is $\overline{L} - NL$ times a. Equating these two quantities yields

$$a = \frac{NLb}{\overline{L} - NL}. \tag{10.36}$$

Equation (10.36) implies $a + b = \overline{L}b/(\overline{L} - NL)$. Substituting this into (10.35) yields

$$w = \bar{e} + \left(\rho + \frac{\overline{L}}{\overline{L} - NL}b\right)\frac{\bar{e}}{q}. \tag{10.37}$$

Equation (10.37) is the *no-shirking condition*. It shows, as a function of the level of employment, the wage that firms must pay to induce workers to exert effort. When more workers are employed, there are fewer unemployed workers and more workers leaving their jobs; thus it is easier for unemployed workers to find employment. The wage needed to deter shirking is therefore an increasing function of employment. At full employment, unemployed workers find work instantly, and so there is no cost to being fired and thus no wage that can deter shirking. The set of points in (NL,w) space satisfying the no-shirking condition (NSC) is shown in Figure 10.2.

Closing the Model

Firms hire workers up to the point where the marginal product of labor equals the wage. Equation (10.23) implies that when its workers are exerting effort, a firm's flow profits are $F(\bar{e}L) - wL$. Thus the condition for the marginal product of labor to equal the wage is

$$\bar{e}F'(\bar{e}L) = w. \tag{10.38}$$

[12] We are assuming that the economy is large enough that although the breakup of any individual job is random, aggregate breakups are not.

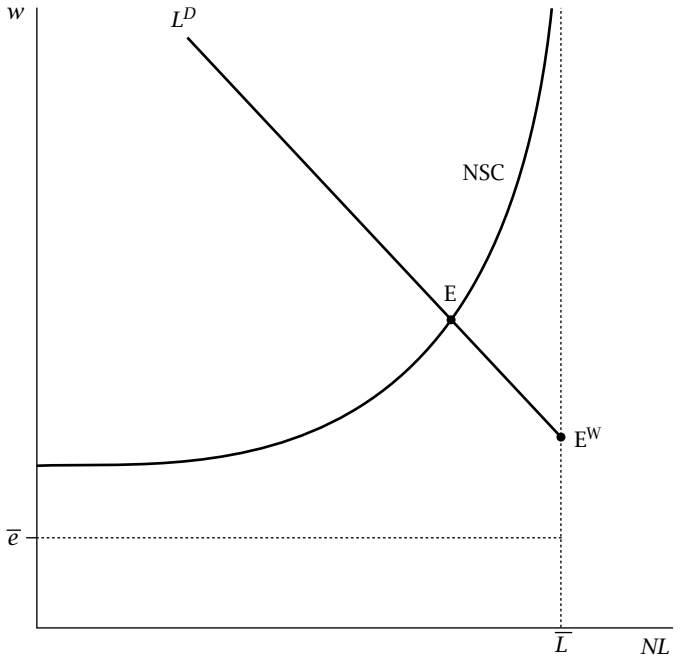


FIGURE 10.2 The Shapiro–Stiglitz model

The set of points satisfying (10.38) (which is simply a conventional labor demand curve) is also shown in Figure 10.2.

Labor supply is horizontal at $\overline{e}$ up to the number of workers, $\overline{L}$, and then vertical. In the absence of imperfect monitoring, equilibrium occurs at the intersection of labor demand and supply. Our assumption that the marginal product of labor at full employment exceeds the disutility of effort ($F'(\overline{e}\,\overline{L}/N) > 1$) implies that this intersection occurs in the vertical part of the labor supply curve. The Walrasian equilibrium is shown as Point E^W in the diagram.

With imperfect monitoring, equilibrium occurs at the intersection of the labor demand curve (equation [10.38]) and the no-shirking condition (equation [10.37]). This is shown as Point E in the diagram. At the equilibrium, there is unemployment. Unemployed workers strictly prefer to be employed at the prevailing wage and exert effort than to remain unemployed. Nonetheless, they cannot bid the wage down: firms know that if they hire additional workers at slightly less than the prevailing wage, the workers will prefer shirking to exerting effort. Thus the wage does not fall, and the unemployment remains.

Two examples may help to clarify the workings of the model. First, a rise in q—an increase in the probability per unit time that a shirker is detected—shifts the no-shirking locus down and does not affect the labor demand

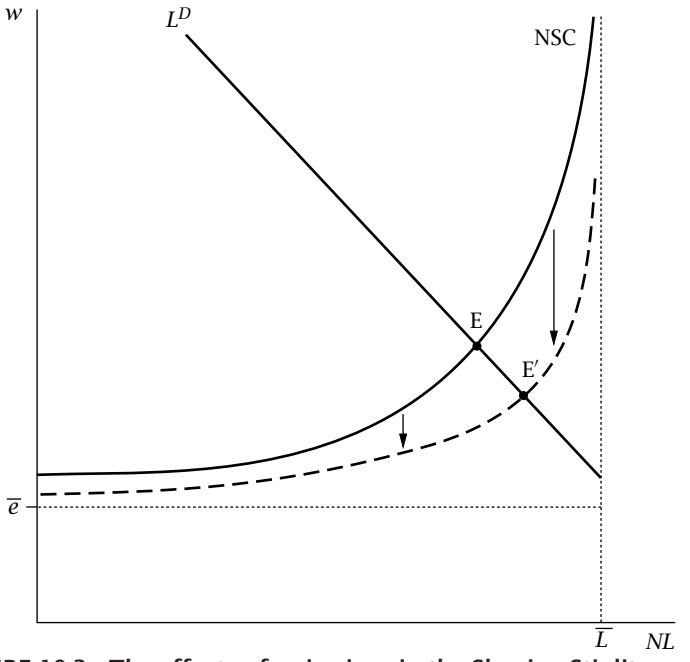


FIGURE 10.3 The effects of a rise in q in the Shapiro–Stiglitz model

curve. This is shown in Figure 10.3. Thus the wage falls and employment rises. As q approaches infinity, the probability that a shirker is detected in any finite length of time approaches 1. As a result, the no-shirking wage approaches $\bar{e}$ for any level of employment less than full employment. Thus the economy approaches the Walrasian equilibrium.

Second, if there is no turnover ($b = 0$), unemployed workers are never hired. As a result, the no-shirking wage is independent of the level of employment. From (10.37), the no-shirking wage in this case is $\bar{e} + \rho\bar{e}/q$. Intuitively, the gain from shirking relative to exerting effort is $\bar{e}$ per unit time. The cost is that there is probability q per unit time of becoming permanently unemployed and thereby losing the discounted surplus from the job, which is $(w - \bar{e})/\rho$. Equating the cost and benefit gives $w = \bar{e} + \rho\bar{e}/q$. This case is shown in Figure 10.4.

Implications

The model implies that there is equilibrium unemployment and suggests various factors that are likely to influence it. Thus the model has some promise as a candidate explanation of unemployment. Unfortunately, the model is so stylized that it is difficult to determine what level of unemployment it predicts.

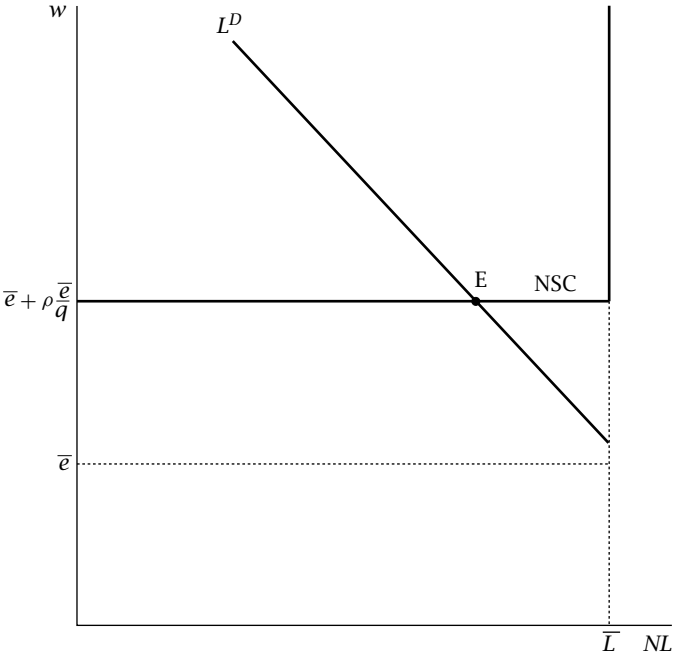


FIGURE 10.4 The Shapiro–Stiglitz model without turnover

With regard to short-run fluctuations, consider the impact of a fall in labor demand, shown in Figure 10.5. w and L move down along the no-shirking locus. Since labor supply is perfectly inelastic, employment necessarily responds more than it would without imperfect monitoring. Thus the model suggests one possible reason that wages may respond less to demand-driven output fluctuations than they would if workers were always on their labor supply curves.[13]

Unfortunately, however, this effect appears to be quantitatively small. When unemployment is lower, a worker who is fired can find a new job more easily, and so the wage needed to prevent shirking is higher; this is the reason the no-shirking locus slopes up. Attempts to calibrate the model suggest that the locus is quite steep at the levels of unemployment we observe. That is, the model implies that the impact of a shift in labor demand

[13] The simple model presented here has the same problem as the simple efficiency-wage model in Section 10.2: it implies that as technological progress continually shifts the labor demand curve up, unemployment trends down. One way to eliminate this prediction is to make the cost of exerting effort, $\overline{e}$, endogenous, and to structure the model so that $\overline{e}$ and output per worker grow at the same rate in the long run. This causes the NSC curve to shift up at the same rate as the labor demand curve in the long run, and thus eliminates the downward trend in unemployment.

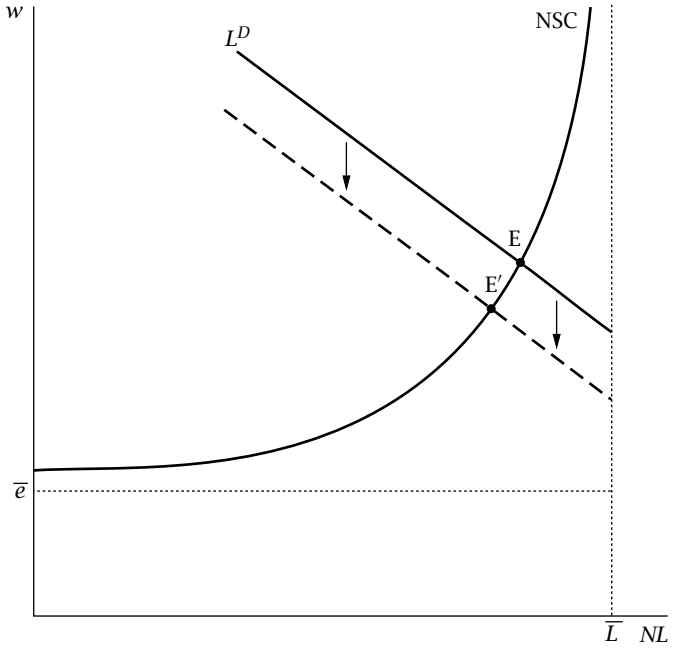


FIGURE 10.5 The effects of a fall in labor demand in the Shapiro–Stiglitz model

falls mainly on wages and relatively little on employment (Gomme, 1999; Alexopoulos, 2004).[14]

Finally, the model implies that the decentralized equilibrium is inefficient. To see this, note that the marginal product of labor at full employment, $\bar{e}F'(\bar{e}\bar{L}/N)$, exceeds the cost to workers of supplying effort, $\bar{e}$. Thus the first-best allocation is for everyone to be employed and exert effort. Of course, the government cannot bring this about simply by dictating that firms move down the labor demand curve until full employment is reached: this policy causes workers to shirk, and thus results in zero output. But Shapiro and Stiglitz note that wage subsidies financed by lump-sum taxes or profits taxes improve welfare. This policy shifts the labor demand curve up, and thus increases the wage and employment along the no-shirking locus. Since the value of the additional output exceeds the opportunity cost of producing it, overall welfare rises. How the gain is divided between workers and firms depends on how the wage subsidies are financed.

[14] In contrast to the simple analysis in the text, these authors analyze the dynamic effects of a shift in labor demand rather than comparing steady states with different levels of demand.

selling occurs when firms require employees to pay a fee when they are hired. If firms are obtaining payments from new workers, their labor demand is higher for a given wage; thus the wage and employment rise as the economy moves up the no-shirking curve. If firms are able to require bonds or sell jobs, they will do so, and unemployment will be eliminated from the model.

Bonding, job selling, and the like may be limited by an absence of perfect capital markets (so that it is difficult for workers to post large bonds, or to pay large fees when they are hired). They may also be limited by workers' fears that the firm may falsely accuse them of shirking and claim the bonds, or dismiss them and keep the job fee. But, as Carmichael (1985) emphasizes, such considerations cannot eliminate these schemes entirely: if workers strictly prefer employment to unemployment, firms can raise their profits by, for example, charging marginally more for jobs. In such situations, jobs are not rationed, but go to those who are willing to pay the most for them. Thus even if these schemes are limited, they still eliminate unemployment. In short, the absence of job fees and performance bonds is a puzzle for the theory.

It is important to keep in mind that the Shapiro–Stiglitz model focuses on one particular source of efficiency wages. Neither its conclusions nor the difficulties it faces in explaining the absence of bonding and job selling are general. For example, suppose firms find high wages attractive because they improve the quality of job applicants on dimensions they cannot observe. Since the attractiveness of a job presumably depends on the overall compensation package, in this case firms have no incentive to adopt schemes such as job selling. Likewise, there is no reason to expect the implications of the Shapiro–Stiglitz model concerning the effects of a shift in labor demand to apply in this case.

As described in Section 10.8, workers' feelings of gratitude, anger, and fairness appear to be important to wage-setting. If these considerations are the reason that the labor market does not clear, again there is no reason to expect the Shapiro–Stiglitz model's implications concerning compensation schemes and the effects of shifts in labor demand to hold. In this case, theory provides little guidance. Generating predictions concerning the determinants of unemployment and the cyclical behavior of the labor market requires more detailed study of the determinants of workers' attitudes and their impact on productivity. Section 10.8 describes some preliminary attempts in this direction.

10.5 Contracting Models

The second departure from Walrasian assumptions about the labor market that we consider is the existence of long-term relationships between firms and workers. Firms do not hire workers afresh each period. Instead, many