Equilibrium Unemployment as a Worker Discipline Device

By CARL SHAPIRO AND JOSEPH E. STIGLITZ*

Involuntary unemployment appears to be a persistent feature of many modern labor markets. The presence of such unemployment raises the question of why wages do not fall to clear labor markets. In this paper we show how the information structure of employer-employee relationships, in particular the inability of employers to costlessly observe workers' on-the-job effort, can explain involuntary unemployment[1] as an equilibrium phenomenon. Indeed, we show that imperfect monitoring necessitates unemployment in equilibrium.

The intuition behind our result is simple. Under the conventional competitive paradigm, in which all workers receive the market wage and there is no unemployment, the worst that can happen to a worker who shirks on the job is that he is fired. Since he can immediately be rehired, however, he pays no penalty for his misdemeanor. With imperfect monitoring and full employment, therefore, workers will choose to shirk.

To induce its workers not to shirk, the firm attempts to pay more than the "going wage"; then, if a worker is caught shirking and is fired, he will pay a penalty. If it pays one firm to raise its wage, however, it will pay all firms to raise their wages. When they all raise their wages, the incentive not to shirk again disappears. But as all firms raise their wages, their demand for labor decreases, and unemployment results. With unemployment, even if all firms pay the same wages, a worker has an incentive not to shirk. For, if he is fired, an individual will not immediately obtain another job. The equilibrium unemployment rate must be sufficiently large that it pays workers to work rather than to take the risk of being caught shirking.

The idea that the threat of firing a worker is a method of discipline is not novel. Guillermo Calvo (1981) studied a static model which involves equilibrium unemployment.[2] No previous studies have treated general market equilibrium with dynamics, however, or studied the welfare properties of such unemployment equilibria. One key contribution of this paper is that the punishment associated with being fired is endogenous, as it depends on the equilibrium rate of unemployment. Our analysis thus goes beyond studies of information and incentives within organizations (such as Armen Alchian and Harold Demsetz, 1972, and the more recent and growing literature on worker-firm relations as a principal-agent problem) to inquire about the equilibrium conditions in markets with these informational features.

The paper closest in spirit to ours is Steven Salop (1979) in which firms reduce turnover costs when they raise wages; here the savings from higher wages are on monitoring costs (or, at the same level of monitoring, from increased output due to increased effort). As in the Salop paper, the unemployment in this paper is definitely involuntary, and not of the standard search theory type (Peter Diamond, 1981, for example). Workers have perfect information about all job opportunities in our model, and unemployed workers strictly prefer to work at wages less than the prevailing market wage (rather than to remain unemployed); there are no vacancies.

*Woodrow Wilson School of Public and International Affairs, and Department of Economics, respectively, Princeton University, Princeton, NJ 08540. We thank Peter Diamond, Gene Grossman, Ed Lazear, Steve Salop, and Mike Veall for helpful comments. Financial support from the National Science Foundation is appreciated.

[1] By involuntary unemployment we mean a situation where an unemployed worker is willing to work for less than the wage received by an equally skilled employed worker, yet no job offers are forthcoming.

[2] In his 1979 paper, Calvo surveyed a variety of models of unemployment, including his hierarchical firm model (also with Stanislaw Wellisz, 1979). There are a number of important differences between that work and this paper, including the specification of the monitoring technology.

The theory we develop has several important implications. First, we show that unemployment benefits (and other welfare benefits) increase the equilibrium unemployment rate, but for a reason quite different from that commonly put forth (i.e., that individuals will have insufficient incentives to search for jobs). In our model, the existence of unemployment benefits reduces the "penalty" associated with being fired. Therefore, to induce workers not to shirk, firms must pay higher wages. These higher wages reduce the demand for labor.

Second, the model explains why wages adjust slowly in the face of aggregate shocks. A decrease in the demand for labor will ultimately cause a lower wage and a higher level of unemployment. In the transition, however, the wage decrease will match the growth in the unemployment pool, which may be a sluggish process.

Third, we show that the market equilibrium which emerges is not, in general, Pareto optimal, where we have taken explicitly into account the costs associated with monitoring. There exist, in other words, interventions in the market that make everyone better off. In particular, we show that there are circumstances in which wage subsidies are desirable. There are also circumstances where the government should intervene in the market by supplying unemployment insurance, even if all firms (rationally) do not. A (small) turnover tax is desirable, because high turnover increases the flow of job vacancies, and hence the flow out of the unemployment pool, making the threat of firing less severe.

Additionally, our theory provides predictions about the characteristics of labor markets which cause the natural rate (i.e., equilibrium level) of unemployment to be relatively high: high rates of labor turnover, high monitoring costs, high discount rates for workers, significant possibilities for workers to vary their effort inputs, or high costs to employers (such as broken machinery) from shirking.

Finally, our theory shows how wage distributions (for identical workers) can persist in equilibrium. Firms which find shirking particularly costly will offer higher wages than other firms do. The dual role wages play by allocating labor and providing incentives for employee effort allows wage dispersion to persist.

Although we have focused our analysis on the labor market, it should be clear that a similar analysis could apply to other markets (for example, product or credit markets) as well. This paper can be viewed as an analysis of a simplified general equilibrium model of an economy in which there are important principal-agent (incentive) problems, and in which the equilibrium entails *quantity constraints* (job rationing). As in all such problems, it is important to identify what is observable, and, based on what is observable, what are the set of feasible contractual arrangements between the parties to the contract. Under certain circumstances, for instance, workers might issue performance bonds and this might alleviate the problems with which we are concerned in this paper. In Section III we discuss the role of alternative incentive devices.

In the highly simplified model upon which we focus here, all workers are identical, all firms are identical, and thus, in equilibrium, all pay the same wage. The assumption that all workers are the same is important, because it implies that being fired carries no stigma (the next potential employer knows that the worker is no more immoral than any other worker; he only infers that the firm for which the worker worked must have paid a wage sufficiently low that it paid the worker to shirk). We have made this assumption because we wished to construct the simplest possible model focussing simply on incentive effects, in which adverse selection considerations play no role. In a sequel, we hope to explore the important interactions between the two fundamental information problems of adverse selection and moral hazard.[3]

The assumption that all firms are the same is not critical for the existence of equilibrium unemployment. Firm heterogeneity will, however, lead to a wage distribution. If the

[3] Other studies have focused on quantity constraints (rationing) with adverse-selection problems. See Stiglitz (1976), Charles Wilson (1980), Andrew Weiss (1980), and Stiglitz and Weiss (1981).

damage that a particular firm incurs as a result of a worker not performing up to standard is larger, the firm will have an incentive to pay the worker a higher wage. Similarly, if the cost of monitoring (detecting shirking) for a firm is large, that firm will also pay a higher wage. Thus, even though workers are all identical, workers for different firms will receive different wages. There is considerable evidence that, in fact, different firms do pay different wages to workers who appear to be quite similar (for example, more capital intensive firms pay higher wages). The theory we develop here may provide part of the explanation of this phenomenon.

In Section I, we present the basic model in which workers are risk neutral. Quit rates and monitoring intensities are exogenous. A welfare analysis of the unemployment equilibrium is provided. In Section II, we comment on extensions of the analysis to situations where monitoring intensities and quit rates are endogenous, and where workers are risk averse. Section III compares the role of unemployment as an incentive device with other methods of enforcing discipline on the labor force.

I. The Basic Model

In this section we formulate a simple model which captures the incentive role of unemployment as described above. Extensions and modifications of this basic model are considered in subsequent sections.

A. *Workers*

There are a fixed number, N, of identical workers, all of whom dislike putting forth effort, but enjoy consuming goods. We write an individual's instantaneous utility function as $U(w, e)$, where w is the wage received and e is the level of effort on the job. For simplicity, we shall assume the utility function is separable; initially, we shall also assume that workers are risk neutral. With suitable normalizations, we can therefore rewrite utility as $U = w - e$. Again, for simplicity, we assume that workers can provide either minimal effort ($e = 0$), or some fixed positive level of $e > 0$.[4] When a worker is unemployed, he receives unemployment benefits of $\bar{w}$ (and $e = 0$).

Each worker is in one of two states at any point in time: employed or unemployed. There is a probability b per unit time that a worker will be separated from his job due to relocation, etc., which will be taken as exogenous. Exogenous separations cause a worker to enter the unemployment pool. Workers maximize the expected present discounted value of utility with a discount rate $r > 0$.[5] The model is set in continuous time.

B. *The Effort Decision of a Worker*

The only choice workers make is the selection of an effort level, which is a discrete choice by assumption. If a worker performs at the customary level of effort for his job, that is, if he does not shirk, he receives a wage of w and will retain his job until exogenous factors cause a separation to occur. If he shirks, there is some probability q (discussed below), per unit time, that he will be caught.[6] If he is caught shirking he will be fired,[7] and forced to enter the unemployment pool. The probability per unit time of acquiring a job while in the unemployment pool (which we call the job acquisition rate, an endogenous variable calculated below) determines the expected length of the unemployment spell he must face. While unemployed he receives unemployment compensation of $\bar{w}$ (also discussed below).

[4] Including effort as a continuous variable would not change the qualitative results.

[5] That is, we assume individuals are infinitely lived, and have a pure rate of time preference of r. They maximize

$$W = E\int_0^\infty u(w(t), e(t))\exp(-rt)\,dt,$$

where we have implicitly assumed that individuals can neither borrow nor lend. Allowing an exponential death rate would not alter the structure of the model; neither would borrowing in the risk-neutral case.

[6] For now we take q as exogenous; later it will be endogenous. The assumption of a Poisson detection technology, like a number of the other assumptions employed in the analysis, is made to ensure that the model has a simple stationary structure.

[7] This will be firm's optimal policy in equilibrium.

The worker selects an effort level to maximize his discounted utility stream. This involves comparison of the utility from shirking with the utility from not shirking, to which we now turn. We define V_E^S as the expected lifetime utility of an employed shirker, V_E^N as the expected lifetime utility of an employed nonshirker, and V_u as the expected lifetime utility of an unemployed individual. The fundamental asset equation for a shirker is given by

$$(1) \qquad rV_E^S = w + (b+q)(V_u - V_E^S),$$

while for a nonshirker, it is

$$(2) \qquad rV_E^N = w - e + b(V_u - V_E^N).$$

Each of these equations is of the form "interest rate times asset value equals flow benefits (dividends) plus expected capital gains (or losses)."[8] Equations (1) and (2) can be solved for V_E^S and V_E^N:

$$(3) \qquad V_E^S = \frac{w + (b+q)V_u}{r+b+q};$$

$$(4) \qquad V_E^N = \frac{(w-e) + bV_u}{r+b}.$$

The worker will choose not to shirk if and only if $V_E^N \geq V_E^S$. We call this the *no-shirking condition* (*NSC*), which, using (3) and (4), can be written as

$$(5) \qquad w \geq rV_u + (r+b+q)e/q \equiv \hat{w}.$$

Alternatively, the *NSC* also takes the form $q(V_E^S - V_u) \geq e$. This highlights the basic implication of the *NSC*: unless there is a penalty associated with being unemployed, everyone will shirk. In other words, if an individual could immediately obtain employment after being fired, $V_u = V_E^S$, and the *NSC* could never be satisfied.

Equation (5) has several natural implications. If the firm pays a sufficiently high wage, then the workers will not shirk. The critical wage, $\hat{w}$, is higher

(a) the higher the required effort (e),

(b) the higher the expected utility associated with being unemployed (V_u),

(c) the lower the probability of being detected shirking (q),

(d) the higher the rate of interest (i.e., the relatively more weight is attached to the short-run gains from shirking (until one is caught) compared to the losses incurred when one is eventually caught),

(e) the higher the exogenous quit rate b (if one is going to have to leave the firm anyway, one might as well cheat on the firm).

C. *Employers*

There are M identical firms, $i = 1, \ldots, M$. Each firm has a production function $Q_i = f(L_i)$, generating an aggregate production function of $Q = F(L)$.[9] Here L_i is firm i's effective labor force; we assume a worker contributes one unit of effective labor if he does not shirk. Otherwise he contributes nothing (this is merely for simplicity). Therefore firms compete in offering wage packages, subject to the constraint that their workers choose not to shirk. We assume that $F'(N) > e$, that is, full employment is efficient.

The monitoring technology (q) is exogenous. Monitoring choices by employers are analyzed in the following section. We assume

[8] A derivation follows: taking V_u as given and looking at a short time interval $[0, t]$ we have

$$V_E = wt + (1-rt)[btV_u + (1-bt)V_E],$$

since there is probability bt of leaving the job during the interval $[0, t]$ and since $e^{-rt} \approx 1 - rt$. Solving for V_E, we have

$$V_E = [wt + (1-rt)btV_u]/[1-(1-rt)(1-bt)].$$

Taking limits as $t \to 0$ gives (1). Equation (2) can be derived similarly.

[9] That is,

$$F(L) \equiv \max_{\{L_i\}} \sum f_i(L_i)$$

such that $\sum L_i = L$. This assumes that in market equilibrium, labor is efficiently allocated, as it will be in the basic model of this section. The modifications required for more general cases, when different firms face different critical no-shirking wages, $\hat{w}_i$, or have different technologies, are straightforward.

that other factors (for example, exogenous noise or the absence of employee specific output measures) prevent monitoring of effort via observing output.

A firm's wage package consists of a wage, w, and a level of unemployment benefits, $\bar{w}$.[10] Each firm finds it optimal to fire shirkers, since the only other punishment, a wage reduction, would simply induce the disciplined worker to shirk again.

It is not difficult to establish that all firms offer the smallest unemployment benefits allowed (say, by law).[11] This follows directly from the *NSC*, equation (5). An individual firm has no incentive to set $\bar{w}$ any higher than necessary. An increase in $\bar{w}$ raises V_u and hence requires a higher w to meet the *NSC*. Therefore, increases in $\bar{w}$ cost the firm both directly (higher unemployment benefits) and indirectly (higher wages). Since the firm has no difficulty attracting labor (in equilibrium), it sets $\bar{w}$ as small as possible. Hence we can interpret $\bar{w}$ in what follows as the minimum legal level, which is offered consistently by all firms.

Having offered the minimum allowable $\bar{w}$, an individual firm pays wages sufficient to induce employee effort, that is, $w = \hat{w}$ to meet the *NSC*. The firm's labor demand is given by equating the marginal product of labor to the cost of hiring an additional employee. This cost consists of wages and future unemployment benefits. For $\bar{w} = 0$,[12] the labor demand is given simply by $f'(L_i) = \hat{w}$, with aggregate labor demand of $F'(L) = \hat{w}$.

[10] More complex employment contracts, for example, wages rising with seniority, are discussed in Section III. With our assumptions of stationarity and identical workers, employers cannot improve on the simple employment provisions considered here.

[11] We are implicitly assuming that the firm cannot offer $\bar{w}$ only to workers who quit. This is so because the firm can always fire a worker who wishes to quit, and it would be optimal for the firm to do so.

[12] For $\bar{w} > 0$ the expected cost of a worker is the wage cost for the expected employment period of $1/b$, followed by $\bar{w}$ for the expected period of unemployment, $1/a$. This generates labor demand given by

$$f'(L_i) = w + \bar{w}b/(a+r).$$

D. *Market Equilibrium*

We now turn to the determination of the equilibrium wage and employment levels. Let us first indicate heuristically the factors which determine the equilibrium wage level.

If wages are very high, workers will value their jobs for two reasons: (a) the high wages themselves, and (b) the correspondingly low level of employment (due to low demand for labor at high wages) which implies long spells of unemployment in the event of losing one's job. In such a situation employers will find they can reduce wages without tempting workers to shirk.

Conversely, if the wage is quite low, workers will be tempted to shirk for two reasons: (a) low wages imply that working is only moderately preferred to unemployment, and (b) high employment levels (at low wages there is a large demand for labor) imply unemployment spells due to being fired will be brief. In such a situation firms will raise their wages to satisfy the *NSC*.

Equilibrium occurs when each firm, taking as given the wages and employment levels at other firms, finds it optimal to offer the going wage rather than a different wage. The key market variable which determines individual firm behavior is V_u, the expected utility of an unemployed worker. We turn now to the calculation of the equilibrium V_u.[13]

The asset equation for V_u, analogous to (1) and (2), is given by

$$rV_u = \bar{w} + a(V_E - V_u), \tag{6}$$

where a is the job acquisition rate and V_E is the expected utility of an employed worker (which equals V_E^N in equilibrium). We can now solve (4) and (6) simultaneously for V_E and V_u to yield

$$rV_E = \frac{(w-e)(a+r)+\bar{w}b}{a+b+r}; \tag{7}$$

$$rV_u = \frac{(w-e)a+\bar{w}(b+r)}{a+b+r}. \tag{8}$$

[13] We have already shown that all firms offer the same employment benefits $\bar{w}$, so V_u is indeed a single number, i.e., an unemployed person's utility is independent of his previous employer.

Substituting the expression for V_u (i.e., (8)) into the *NSC* (5) yields the *aggregate NSC*

$$(9) \qquad w \geq \bar{w} + e + e(a+b+r)/q.$$

Notice that the critical wage for nonshirking is greater: (a) the smaller the detection probability q; (b) the larger the effort e; (c) the higher the quit rate b; (d) the higher the interest rate r; (e) the higher the unemployment benefit ($\bar{w}$); and (f) the higher the flows out of unemployment a.

We commented above on the first four properties; the last two are also unsurprising. If the unemployment benefit is high, the expected utility of an unemployed individual is high, and therefore the punishment associated with being unemployed is low. To induce individuals not to shirk, a higher wage must be paid. If a is the probability of obtaining a job per unit of time, $1/a$ is the expected duration of being unemployed. The longer the duration, the greater the punishment associated with being unemployed, and hence the smaller the wage that is required to induce nonshirking.

The rate a itself can be related to more fundamental parameters of the model, in a steady-state equilibrium. In steady state the flow *into* the unemployment pool is bL where L is aggregate employment. The flow *out* is $a(N-L)$ (per unit time) where N is the total labor supply. These must be equal, so $bL = a(N-L)$, or

$$(10) \qquad a = bL/(N-L).$$

Substituting for a into (9), the aggregate *NSC*, we have

$$(11) \quad w \geq e + \bar{w} + \frac{e}{q}\left(\frac{bN}{(N-L)} + r\right)$$

$$= e + \bar{w} + (e/q)(b/u + r) \equiv \hat{w},$$

where $u = (N-L)/N$, the unemployment rate. This constraint, the aggregate *NSC*, is graphed in Figure 1. It is immediately evident that *no shirking is inconsistent with full employment*. If $L = N$, $a = +\infty$, so any shirking worker would immediately be rehired. Knowing this, workers will choose to shirk.

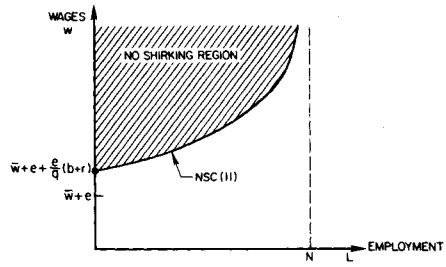

FIGURE 1. THE AGGREGATE NO-SHIRKING CONSTRAINT

The equilibrium wage and employment level are now easy to identify. Each (small) firm, taking the aggregate job acquisition rate a as given, finds that it must offer at least the wage $\hat{w}$. The firm's demand for labor then determines how many workers are hired at the wage. Equilibrium occurs where the aggregate demand for labor intersects the aggregate *NSC*. For $\bar{w} = 0$, equilibrium occurs when

$$F'(L) = e + (e/q)(bN/(N-L) + r).$$

The equilibrium is depicted in Figure 2.[14] It is important to understand the forces which cause E to be an equilibrium. From the firm's point of view, there is no point in raising wages since workers are providing effort and the firm can get all the labor it wants at w^*. Lowering wages, on the other hand, would induce shirking and be a losing idea.[15]

From the worker's point of view, *unemployment is involuntary*: those without jobs would be happy to work at w^* or lower, but cannot make a credible promise not to shirk at such wages.

[14] Aggregate labor demand is $F'(L)$ only when $\bar{w} = 0$ (see fn. 12).

[15] We have assumed that output is zero when an individual shirks, but we need only assume that a shirker's output is sufficiently low that hiring shirking workers is unprofitable.

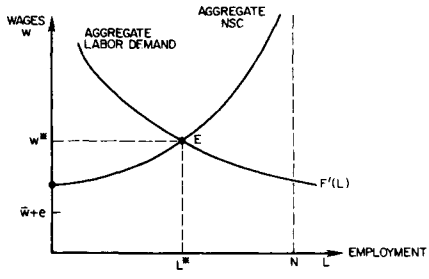

FIGURE 2. EQUILIBRIUM UNEMPLOYMENT

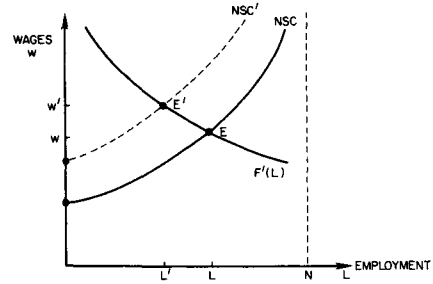

FIGURE 3. COMPARATIVE STATICS

Note: A decrease in the monitoring intensity q, or an increase in the quit rate b, leads to higher wages and more unemployment

Notice that the type of unemployment we have characterized here is very different from search unemployment. Here, all workers and all firms are identical. There is perfect information about job availability. There is a different information problem: firms are assumed (quite reasonably, in our view) not to be able to monitor the activities of their employees costlessly and perfectly.

E. *Simple Comparative Statics*

The effect of changing various parameters of the problem may easily be determined. As noted above, increasing the quit rate b, or decreasing the monitoring intensity q, decreases incentives to exert effort. Therefore, these changes require an increase in the wage necessary (at each level of employment) to induce individuals to work, that is, they shift the *NSC* curve upwards (see Figure 3). On the other hand, they leave the demand curve for labor unchanged, and hence the equilibrium level of unemployment and the equilibrium wage are both increased. Increases in unemployment benefits have the same impact on the *NSC* curve, but they also reduce labor demand as workers become more expensive, so they cause unemployment to rise for two reasons.

Inward shifts in the labor demand schedule create more unemployment. Due to the *NSC*, wages cannot fall enough to compensate for the decreased labor demand. The transition to the higher unemployment equilibrium will not be immediate: wage decreases by individual firms will only become attractive as the unemployment pool grows. This provides an explanation of wage sluggishness.

F. *Welfare Analysis*

In this section we study the welfare properties of the unemployment equilibrium. We demonstrate that the equilibrium is not in general Pareto optimal, when information costs are explicitly accounted for.

We begin with the case where the owners of the firms are the same individuals as the workers, and ownership is equally distributed among N workers. The central planning problem is to maximize the expected utility of the representative worker subject to the *NSC* and the resource constraint:

$$(12)\quad \max_{w,\bar{w},L}\ (w-e)L+\bar{w}(N-L)$$

$$\text{subject to } w \geq e+\bar{w}+(e/q)((bN/(N-L))+r) \quad (NSC)$$

$$\text{subject to } wL+\bar{w}(N-L) \leq F(L) \quad \text{(Feasibility)}$$

$$\text{subject to } \bar{w} \geq 0.$$

Since workers are risk neutral it is easy to check[16] that the optimum involves $\bar{w}$ at the minimum allowable level, which is assumed to be 0. The reason is that increases in $\bar{w}$ tighten the *NSC*, so all payments should be made in the form of w rather $\bar{w}$.

Setting $\bar{w} = 0$, the problem simplifies to

$$(12') \quad \max_{w,L} (w - e)L$$

subject to $w \geq e + (e/q)((bN/(N-L)) + r)$;

and $wL \leq F(L)$.

The set of points which satisfy the constraints is shaded in Figure 4. Iso-utility curves are rectangular hyperboles. So long as $F'(L) > e$, these are steeper than the average product locus, so the optimum occurs at point A where the *NSC* intersects the curve $w = F(L)/L$, that is, where wages equal the average product of labor. In contrast, the market equilibrium occurs at E where the marginal product of labor curve, $w = F'(L)$, intersects the *NSC* (Figure 2). Observe that in the case of constant returns to scale, $F'(L)L = F(L)$, so the equilibrium is optimal.

Wages should be subsidized, using whatever (pure) profits can be taxed away. An equivalent way to view the social optimum is a tax on unemployment to reduce shirking incentives; the wealth constraint on the unemployed requires that $\bar{w} \geq 0$, or equivalently that profits after taxes be nonnegative.[17] The optimum can be achieved by taxing away all profits and financing a wage subsidy of τ, shown in Figure 4. *The "natural" unemployment rate is too high.*

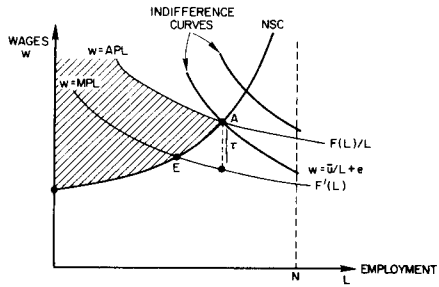

FIGURE 4. SOCIAL OPTIMUM AT A

In the case where the workers and the owners are distinct individuals, the tax policy described above would reduce profits, increase wages, and increase employment levels. While it would increase aggregate output (net of effort costs), such a tax policy would *not* constitute a Pareto improvement, since profits would fall. For this reason, the equilibrium is Pareto optimal in this case, even though it fails to maximize net national product. We thus have the unusual result that the Pareto optimality of the equilibrium depends upon the distribution of wealth. The standard separation between efficiency and income distribution does not carry over to this model.

It should not be surprising that the equilibrium level of unemployment is in general inefficient. Each firm tends to employ too few workers, since it sees the private cost of an additional worker as w, while the social cost is only e, which is lower. On the other hand, when a firm hires one more worker, it fails to take account of the effect this has on V_u (by reducing the size of the unemployment pool). This effect, a negative externality imposed by one firm on others as it raises its

[16] Formally,

$$\mathcal{L} = (w - e)L + \bar{w}(N - L) + \lambda[w - e - \bar{w} - (e/q)(bN/(N - l) + r)] + \mu[F(L) - wL - \bar{w}(N - L)].$$

Differentiating with respect of w and $\bar{w}$ yields

$$\mathcal{L}_w = L + \lambda - \mu L \leq 0 \text{ and } = 0 \text{ if } w > 0.$$

$$\mathcal{L}_{\bar{w}} = (N - L) - \lambda - \mu(N - L) \leq 0 \text{ and } = 0 \text{ if } \bar{w} > 0.$$

We know $w > 0$ by the *NSC*, so $\mathcal{L}_w = 0$, i.e., $L(1 - \mu) + \lambda = 0$. Therefore, since $\lambda > 0$, $\mu > 1$. But then $\mathcal{L}_{\bar{w}} = (N - L)(1 - \mu) - \lambda < 0$. This implies that $\bar{w} = 0$.

[17] The constraint $\bar{w} \geq 0$ can be rewritten, using the resource constraint, as $F(L) - wL \geq 0$, i.e., $\pi \geq 0$.

level of employment, tends to lead to overemployment. In the simple model presented so far, the former effect dominates, and the natural level of unemployment is too high. This will not be true in more general models, however, as we shall see below.

II. Extensions

In this section we describe how the results derived above are modified or extended when we relax some of the simplifying assumptions. We discuss three extensions in turn: endogenous monitoring, risk aversion, and endogenous turnover. Detailed derivations of the claims made below are available in our earlier working paper.

A. *Endogenous Monitoring*

When employees can select the monitoring intensity q, they can trade off stricter monitoring (at a cost) with higher wages as methods of worker discipline. In general, firms' monitoring intensities will not be optimal, due to the externalities between firms described above. In general, it is not possible to ascertain whether the equilibrium entails too much or too little employment. In the case of constant returns to scale ($F(L) = L$), however (which led to efficiency with exogenous monitoring), the competitive equilibrium involves too much monitoring and too much employment.

The result is not as unintuitive as it first seems: each firm believes that the only instrument at its control for reducing shirking is to increase monitoring. There is, however, a second instrument: by reducing employment, workers are induced not to shirk. This enables society to save resources on monitoring (supervision). These gains more than offset the loss from the reduced employment.

It is straightforward to see how this policy may be implemented. If firms can be induced to reduce their monitoring, welfare will be increased. Hence a tax on monitoring, with the proceeds distributed, say, as a lump sum transfer to firms, will leave the no-shirking constraint/national-resource constraint unaffected, but will reduce monitoring.

B. *Risk Aversion*

With risk neutrality, the optimum and the market both involve $\bar{w} = 0$. Clearly $\bar{w} = 0$ cannot be optimal if workers are highly risk averse and may be separated from their jobs for exogenous reasons. Yet the market always provides $\bar{w} = 0$ (or the legal minimum). The proof above that $\bar{w} = 0$ carries over to the case of risk-averse workers.

When equilibrium involves unemployment, firms have no difficulty attracting workers and hence offer $\bar{w} = 0$, since $\bar{w} > 0$ merely reduces the penalty of being fired. When *other* firms offer $\bar{w} = 0$, this argument is only strengthened: unemployed workers are even easier to attract. It is striking that the market provides no unemployment benefits even when workers are highly risk averse. Clearly the social optimum involves $\bar{w} > 0$ if risk aversion is great enough. This may provide a justification for mandatory minimum benefit levels.

C. *Endogenous Turnover*

In general a firm's employment package will influence the turnover rate it experiences among its employees. Since the turnover rate b affects the rate of hiring out of the unemployment pool, and hence V_u, it affects other firms' no-shirking constraints. Because of this externality, firms' choices of employment packages will not in general be optimal. This type of externality is similar to search externalities in which, for example, one searcher's expected utility depends on the number or mix of searchers remaining in the market. In the current model, policies which discourage labor turnover are attractive as they make unemployment more costly to shirkers.

III. Alternative Methods for the Enforcement of Discipline

This paper has explored a particular mechanism for the enforcement of discipline: individuals who are detected shirking are fired, and in equilibrium the level of unemployment is sufficiently large that this threat serves as an effective deterrent to shirking. The

question naturally arises whether there are alternative, less costly, or more effective discipline mechanisms.

A. *Performance Bonds*

The most direct mechanism by which discipline might be enforced is through the posting by workers of performance bonds. Under this arrangement the worker would forfeit the bond if the firm detected him shirking. One problem with this solution is that workers may not have the wealth to post bond.[18] A more fundamental problem with this mechanism is that the firm would have an incentive to *claim* that the worker shirked so that it could appropriate the bond. Assuming, quite realistically, that third parties cannot easily observe workers' effort (indeed, it is usually more costly for outsiders to observe worker inputs than for the employer to do so), there is no simple way to discipline the *firm* from this type of opportunism.

Having recognized this basic point, it is easy to see that a number of other plausible solutions face the same difficulty. For example, consider an employment package which rewards effort by raising wages over time for workers who have not been found shirking. This is in fact equivalent to giving the worker a level wage stream, but taking back part of his earlier payments as a bond, which is returned to him later. Therefore, by the above argument, the firm will have an incentive to fire the worker when he is about to enter the "payoff" period in which he recovers his bond. This is the equivalent to the firm's simply appropriating the bond. It is optimal for the firm to replace expensive senior workers by inexpensive junior ones.[19]

Clearly the firm's reputation as an honest employer can partially solve this problem; the employer is implicitly penalized for firing a worker if this renders him less attractive to prospective employees. Yet this reputation mechanism may not work especially well, since prospective employees often do not know the employer's record, and previous dismissals may have been legitimate (it is not possible for prospective employees to distinguish legitimate from unfair earlier dismissals, if they are aware of them at all). If the reputation mechanism is less than perfect, it will be augmented by the unemployment mechanism.

B. *Other Costs of Dismissal*

Unemployment in the model above serves the role of imposing costs on dismissed workers. If other costs of dismissal are sufficiently high, workers may have an incentive to exert effort even under conditions of full employment. Examples of such costs are search costs, moving expenses, loss of job-specific human capital, etc. In markets where these costs are substantial, the role of equilibrium unemployment is substantially diminished. The effect we have identified above will still be present, however, when effort levels are continuous variables: each firm will still find that employee effort is increasing with wages, so wages will be bid up somewhat above their full-employment level. The theory predicts that involuntary (as well as frictional) unemployment rates will be higher for classes of workers who have lower job switching costs.

[18] This is especially true if detection is difficult (low q) so that an effective bond must be quite large. Even if workers could borrow to post the bond, so long as bankruptcy is possible, the incentives for avoiding defaulting on the bond are not different from the incentives to avoid being caught shirking by the firm in the absence of a bond. Note once again the importance of the wealth distribution in determining the nature of the equilibrium. If all individuals inherit a large amount of wealth, then they could post bonds.

[19] In competitive equilibrium, the average (discounted) value of the wage must be equal to the average (discounted) value of the marginal product of the worker. If there is a bonus for not shirking, *initially* the wage must be below the value of the marginal product. It is as if the worker were posting a bond (the difference between his marginal product and the wage), and as such this scheme is susceptible to precisely the same objections raised against posting performance bondings. The employer has an incentive to appropriate the bond. Since workers know this, this is not a viable incentive scheme. For a fine study in which firms' reputations are assumed to function so as to make this scheme viable, see Edward Lazear (1981).

C. *Heterogeneous Workers*

The strongest assumption we have made is that of identical workers. This assumption ruled out the possibility that firing a worker would carry any stigma. Such a stigma could serve as a discipline device, even with full employment.[20] In reality, of course, employers *do* make wage offers which are contingent on employment history. Such policies make sense when firms face problems of adverse selection.

We recognize that workers' concern about protecting their reputations as effective, diligent workers may provide an effective incentive for a disciplined labor force.[21] Shapiro's earlier (1983) analysis of reputation in product markets showed, however, that for reputations to be an effective incentive device, there must be a cost to the loss of reputation. It is our conjecture that, under plausible conditions, even when reputations are important, equilibrium will entail some use of unemployment as a discipline device for the labor force, at least for lower-quality workers. An important line of research is the study of labor markets in which adverse selection as well as moral hazard problems are present. In this context, our model should provide a useful complement to the more common studies of adverse selection in labor markets.

IV. Conclusions

This paper has explored the role of unemployment, or job rationing, as an incentive device. We have argued that when it is costly to monitor individuals, competitive equilibrium will be characterized by unemployment, but that the natural rate of unemployment so engendered will not in general be optimal. We have identified several forces at work, some which tend to make the market equilibrium unemployment rate too high, and others which tend to make it too small. Each firm fails to take into account the consequences of its actions on the level of monitoring and wages which other firms must undertake in order to avoid shirking by workers. Although these externalities are much like pecuniary externalities, they are important, even in economies with a large number of firms.[22] As a result, we have argued that there is scope for government interventions, both with respect to unemployment benefits and taxes or subsidies on monitoring and labor turnover, which can (if appropriately designed) lead to Pareto improvements.

The type of unemployment studied here is not the only or even the most important source of unemployment in practice. We believe it is, however, a significant factor in the observed level of unemployment, especially in lower-paid, lower-skilled, blue-collar occupations. It may well be more important than frictional or search unemployment in many labor markets.

[20] See Bruce Greenwald (1979) for a simple model in which those who are in the "used labor market" are in fact a lower quality than those in the "new" labor market.

[21] This suggests once again that our results may be most significant in labor markets for lower-quality workers: in such markets employment histories are utilized less and workers already labeled as below average in quality have less to lose from being labeled as such.

[22] For a more general discussion of pecuniary, or more general market mediated externalities, with applications to economies with important adverse selection and moral hazard problems, see Greenwald and Stiglitz (1982).

REFERENCES

Alchian, Armen A. and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Calvo, Guillermo A., "Quasi-Walrasian Theories of Unemployment," *American Economic Review Proceedings*, May 1979, *69*, 102–06.

_____, "On the Inefficiency of Unemployment," Columbia University, October 1981.

_____ **and Wellisz, Stanislaw,** "Hierarchy, Ability and Income Distribution," *Journal of Political Economy*, October 1979, *87*, 991–1010.

Diamond, Peter, "Mobility Costs, Frictional Unemployment, and Efficiency," *Journal*

of Political Economy, August 1981, *89*, 798–812.

Greenwald, Bruce, C. N., *Adverse Selection in the Labor Market*, New York; London: Garland, 1979.

_____ **and Stiglitz, Joseph E.,** "Pecuniary Externalities," unpublished, Princeton University, 1982.

Lazear, Edward P., "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, *71*, 606–20.

Salop, Steven C., "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Shapiro, Carl, "Premiums for High Quality Products as Returns to Reputations," *Quarterly Journal of Economics*, November 1983, *98*, 658–79.

_____ **and Stiglitz, Joseph E.,** "Equilibrium Unemployment as a Worker Discipline Device," Discussion Papers in Economics, No. 28, Woodrow Wilson School, Princeton University, April 1982.

Stiglitz, Joseph E., "Prices and Queues as Screening Devices in Competitive Markets," IMSSS Technical Report No. 212, Stanford University, 1976.

_____ **and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

Wilson, Charles, "The Nature of Equilibrium in Markets with Adverse Selection," *Bell Journal of Economics*, Spring 1980, *11*, 108–30.