Motivation and Markets[1]

W. Bentley MacLeod
Boston College and C.R.D.E.

James M. Malcomson
University of Southampton, U.K.

Previous version: August 1994
This version: March 1996

[1] The authors would like to thank Daniel Parent for excellent research assistance and, for helpful comments, Paul Beaudry, Lorne Carmichael, Janet Currie, Greg Dow, Avner Grief, Bengt Holmström, Stephen Jones, Ashok Kotwal, Thomas Lemieux, and participants in the European Summer Symposium on Economic Theory 1993 (especially David Canning, Bob Gibbons, Dilip Mookherjee and Michael Riordan) held at the Studienzentrum Gerzensee of the Swiss National Bank. The financial support of the Social Science and Humanities Research Council of Canada and the UK Economic and Social Research Council (ESRC) is gratefully acknowledged. The work was funded by ESRC award number R000231484.

Abstract

In standard shirking models of efficiency wages, workers are motivated only by high wages. Yet 23% of young US workers report receiving some form of performance pay. This paper extends the efficiency wage framework using the theory of self-enforcing agreements to allow for performance pay in the form of bonuses. The result is a simple model of wage formation that helps explain a number of apparently unrelated phenomena in labor markets. First, in efficient markets performance pay is preferred to an efficiency wage when the cost of having a job vacant is low and qualified workers are in short supply. Second, more capital intensive industries offer higher pay than less capital intensive industries, as observed in studies of inter-industry wages differentials. Third, sustaining an efficient outcome requires a social convention similar to the notion of a fair wage, although the outcome itself is determined by fundamentals and not by exogenously imposed notions of what is fair. Finally, a two-sector version of the model makes some predictions about the relationships between turnover and wages and between wages, growth and unemployment.

JEL classifications: D82, J41, O15.
Keywords: efficiency wages, merit pay, migration, motivation, wage differentials.

1 Introduction

When individual performance is not easily measured in an objective way, there are limitations to the use of legally enforceable performance pay to motivate employees. This paper shows that the use of self-enforcing agreements as an alternative can help us understand a number of apparently unrelated phenomena in labor markets. The model developed here has implications that go beyond the Shapiro and Stiglitz (1984) shirking model of efficiency wages commonly applied to such situations. It makes predictions about the form of efficient incentive agreements. It provides a reason for more capital intensive jobs to offer higher pay. It also provides a reason why notions of fairness may play an important role in the minds of labor market participants without appearing in econometric studies as an important exogenous determinant of labor market outcomes.

Efficiency wage models rely on high wages to provide incentives, yet this is not the only form of compensation we observe in practice. Firms also motivate workers with promises of bonuses and promotions. In many cases these promises are only informal or are based on subjective assessments of performance that make them unenforceable at law. For example, in the 1990 Workplace Industrial Relations Survey for the UK, 34% of employees were reported as receiving some form of *merit pay*, "which depended on a subjective judgement by a supervisor or manager of the individual's performance" (Millward, Stevens, Smart and Hawes (1992, p. 258)). In the 1988-90 National Longitudinal Survey of Youth (NLSY) for the US, 23% of the individuals reported receiving some form of performance pay, though we do not know the extent to which this pay is based on subjective assessments. Moreover, employees paid fixed wages or salaries may get promoted for good performance as subjectively assessed, so figures for merit pay understate the percentage of employees who are rewarded on this basis.

MacLeod and Malcomson (1989; 1993) show that, provided the rent from the employment relationship is sufficiently large, firms may use either efficiency wages (high wages combined with the threat of dismissal) or performance pay (workers paid an end-of-period bonus if they perform well) to motivate employees. This rent is the difference between the returns to the current relationship and those in the spot market for labor.[1] Should either the worker shirk or the firm not pay the bonus, a separation occurs that harms the offending party through a loss of the future rent from the relationship. A practical example of this is the case of the highly paid traders discussed in Stewart (1993) who quit First Boston Bank because their bonuses were smaller than they felt they had been promised, despite receiving total compensation of over one-half million dollars. When there are sufficient rents to the relationship to enforce an agreement in this way, we say the agreement is *self-enforcing*.

This paper models the choice between pay for performance and efficiency wages to provide motivation. The amount of rent needed to make either form of compensation

[1] In the model, firms and workers are risk neutral and hence we can write the total return to a relationship as the sum of profit plus utility. The *rent* to the relationship is the difference between this sum and the sum of profit plus utility if firm and worker were to take their next best market alternatives.

self-enforcing is the same, hence the choice between them does not depend solely on the characteristics of the employment relationship. In the spirit of Hayek (1982), we examine the form of compensation that an efficient market would select. It is the role of the market that distinguishes the present paper from MacLeod and Malcomson (1989; 1993).

With some jobs, the cost of having the position vacant is high relative to the value of a worker's time. Examples are jobs in capital intensive industries where workers are either on an assembly line or involved in surveying the operation of expensive equipment. If such a position is not filled, the equipment may have to be closed down or else used only inefficiently. At an efficient equilibrium firms must, therefore, always be able to replace workers quickly. But if a firm can replace a worker immediately with a close substitute on similar terms, it cannot be earning a rent from employing that worker. In this case, the rent required to make the compensation agreement self-enforcing must go to the worker in the form of a high wage, with the threat of job loss to provide motivation. A rent must also go to new employees. If it does not, the firm could increase its current profits by replacing a current employee with a new one at a lower wage. Anticipating this, the current employee would expect no rent from future employment and so shirk. In an efficient equilibrium, therefore, firms pay efficiency wages.

Conversely, consider jobs for which workers are in short supply, for example sports stars or traders in financial markets. Though these workers earn high wages, they could earn high wages at other jobs and hence the threat of firing is not a motivator. In this case the efficient market equilibrium can be achieved with performance pay. Where performance is subjective so that the performance pay is not enforceable in court, firms have an incentive to renege on a bonus at the end of the period. Performance pay can nevertheless be enforced by the threat of workers either quitting or performing poorly in the future, but only if the firm receives a future rent from continued employment. The actions of the disgruntled traders at First Boston described above show that such a threat is a practical possibility. In this case, the rent from hiring new employees must also go to firms as otherwise employees would quit for higher pay even if paid the bonus.

We also consider a simple two sector economy in which both efficiency wages and performance pay occur in equilibrium. The model predicts that workers in capital intensive jobs receive higher pay, consistent with the evidence found in Dickens and Katz (1987) and Troske (1994). There is then an ambiguous relationship between turnover and pay. In industries that use high wages to motivate workers, relative wages increase with the rate of turnover. Conversely, when workers are in short supply and incentives are provided with bonus payments, turnover is negatively related to total compensation. In the absence of a motivation problem, turnover would not effect equilibrium wages in this model.

The equilibrium in this two sector economy exhibits some interesting dynamics. We consider the characteristics of the equilibrium when there is a spurt of growth in the capital intensive sector, resulting in labor moving into that sector and thus increasing turnover in the less capital intensive sector. This temporary increase in turnover raises the rent required for an agreement to be self-enforcing, thus resulting in a short run increase in unemployment and inequality, and an initial slowdown in total output. All of these effects are temporary, but they signal that the benefits of economic growth to workers may be only long run and not immediate.

The model also provides a way to think about the role of fairness in wage determination. Survey evidence in Kaufman (1984), **?**, Blinder and Choi (1990) and Bewley (1993) indicates that firms believe that treating workers fairly is important. Of the managers interviewed by Blinder and Choi (1990), 79% perceived "a wage reduction to take advantage of labor market slack" as unfair, 95% thought their workers would perceive it as unfair. Moreover, 95% also thought that a reputation for an unfair wage policy would result in reduced work effort. This illustrates how deeply the convention is ingrained on *both* sides of the labor market. Kaufman (1984) comes to a similar conclusion for Britain, even for non-unionized firms.

In the current model, the level of wages and form of compensation depend on market conditions, turnover rates and vacancy costs, and not on any exogenously imposed notions of fairness. However, the existence of a self-enforcing agreement that sustains equilibrium depends on a convention such as fairness. When employees are paid efficiency wages (thus receiving the rent from employment), firms have an incentive to try to lower wages for the current period because it is high future wages, not high current wages, that provide the incentive to perform. If employees anticipate their succeeding, the incentive to perform in the previous period is removed. Sustaining an efficient equilibrium thus requires a convention such as fairness that prevents wages being bid down. In the model this convention is an equilibrium because firms believe wage cuts today will result in workers shirking *in the future*, in exactly the manner described in the survey evidence. Conversely, when bonus payments are used, firms are induced to pay the bonus only because of a rent from continuing the employment in the future. Once a bonus has been paid, workers have an incentive to try to raise wages for the current period because it is only the future rent that provides the incentive for firms to pay. But if firms anticipate their succeeding, the incentive to pay the bonus in the previous period is destroyed. Thus sustaining an efficient equilibrium in this case requires a convention that prevents wages being bid up.

The agenda of the paper is as follows. Section 2 reviews the theory of self-enforcing agreements and how the existence of such agreements is related to the concept of a fair wage. Section 3 examines the nature of a market equilibrium in self-enforcing agreements. We illustrate that the form of agreement is related to the cost of having a job vacant. In Section 4 we consider a two sector version of the model. We examine the form of inter-industry wage variation as a function of capital intensity in the two sectors in Section 4.1. Section 4.2 presents some simulation results of the effect of growth in the capital intensive sector.

2 Self-Enforcing Agreements

We model a long term employment relationship in the following way. At the beginning of each period worker and firm agree on a compensation package that includes a fixed wage component and possibly an end of period bonus. The worker then decides the level of effort to supply, followed by the firm's decision on the level of bonus pay. Finally, either party may decide to leave the current relationship. In the spirit of the incomplete contracts literature we suppose that, though the performance of the worker is observed by the firm, important aspects of performance are not measurable in a way that can be verified in court. Hence the bonus cannot be made a legally binding contract tying pay to performance.[2] As Posner (1986, p. 79) observes, rather than try to enforce performance with the intervention of the courts, it may be more efficient to use an agreement that is *self-enforcing.* An agreement is made self-enforcing by the worker quitting or shirking in the future, and by the firm firing the worker or cutting the bonus, if the agreement is broken. This section provides a complete characterization of the set of self-enforcing agreements and illustrates the role of a fair wage convention in ensuring the existence of an equilibrium.

Consider a market for homogeneous labor with discrete periods that correspond to the minimum time a worker is employed. There are L identical workers and, in contrast to MacLeod and Malcomson (1989; 1993), free entry of firms with identical jobs. If there are J_t jobs at time t, the sunk cost of creating an additional job vacancy at t is continuous and non-decreasing in J_t. We denote this cost by $c_t(J_t)$. It consists of the sunk cost element of the capital equipment required for the additional employee and any costs of reorganizing the production process. This cost must be incurred before the matching process starts for period t so that, at the time matching takes place, J_t is fixed.

Employed workers get utility $W_t - ve_t$ from being paid W_t and exerting effort e_t in period t. For simplicity, we assume workers either work ($e_t = 1$) or shirk ($e_t = 0$), with $v > 0$ the disutility of work. The model is more general than this unidimensional interpretation suggests. Not shirking can correspond to the production of a vector of quantitative and qualitative dimensions of performance at an acceptable level. The profit from a job employing a worker in period t is $pe_t - W_t$, where p is the revenue product of labor. Each firm individually takes p as given but it may in aggregate be a decreasing function of the total employment of non-shirking workers in all firms because of a downward sloping demand curve for the product.[3] Workers without jobs receive utility $u_t > 0$ per period. Vacant jobs and jobs with shirking workers earn zero profits, so firms never retain a worker who always shirks.

At the beginning of period t, unemployed workers and firms with vacancies meet in a market to negotiate employment. Firms make offers that workers can either accept or reject. If an offer is rejected, worker and firm re-enter the market with expected future utility $\bar{U}_t$ and profits $\bar{\Pi}_t$, respectively. The pay for period t consists

[2] MacLeod (1996) provides a formal model that links this contract incompleteness to job complexity as measured by the number of different tasks that a worker is called upon to carry out.

[3] This could also be because of external decreasing returns to scale, in which case the welfare comparisons discussed later need minor amendments.

of a base wage w_t that the firm pays whatever the worker's performance and a bonus b_t that the firm agrees to pay as long as the worker does not shirk. Consistent with employment at will, the firm is committed to the wage for the current period of employment only. However, for modelling purposes we assume that, at the time of hiring, worker and firm reach agreement (possibly implicit) on wages and bonuses for the life of the relationship. Below we discuss sufficient conditions that ensure that the firm does not renege on these future wage payments.

The wage w_t can be paid either at the start of the period or, as long as the courts enforce it, at the end. (In principle, w_t might be negative, in which case it would be a bond the worker pays the firm in period t, but this is never required for any of the equilibria we discuss below.) The bonus b_t may be simply a periodic bonus or take the form of a piece rate, commission on sales, crop share, promotion to a higher paying grade, or any other type of performance related pay based on performance that is not verifiable in court. It thus differs from the fixed wage w_t in that it is paid only at the end of the period and so can be made conditional on the performance of the worker in period t, but can never be legally enforced because courts cannot verify performance.

Once employment for the current period has been agreed, the worker chooses whether to work or to shirk. The firm observes the worker's performance and then decides whether to pay the bonus b_t. At the end of the period, the worker and the firm decide whether to stay matched for the next period. There is a positive probability that they will separate at this point because it becomes unprofitable to continue the job for purely exogenous reasons unrelated to pay. The timing of events is shown in figure 1.

If neither party reneges, the wage payment in period t is $W_t = w_t + b_t$. Let ρ_t denote the probability that a match at $t-1$ continues at t if neither party reneges. We assume that ρ_t is independent of how long the match has lasted and, for the moment, that it is less than one only because some jobs become unprofitable for exogenous reasons. For any wage profile, the expected future utility of a worker employed in period t in a match starting at τ if both parties abide by the agreement is

$$U_t = W_t - v + \delta \left[\rho_{t+1} U_{t+1} + (1 - \rho_{t+1}) \bar{U}_{t+1} \right], \text{ for all } t \geq \tau, \tag{1}$$

where $\delta \in (0, 1)$ is the discount factor and $\bar{U}_{t+1}$ the expected future utility of an employee whose match terminates at the end of period t.[4] The firm's expected future profit from employing the worker if both abide by the agreement is

$$\Pi_t = p - W_t + \delta \rho_{t+1} \Pi_{t+1}, \text{ for all } t \geq \tau. \tag{2}$$

(If a job becomes unprofitable for exogenous reasons, the profit from that job is zero thereafter.)

To be self-enforcing, an agreement must be at least as good for both worker and firm as they could get elsewhere in the market. The payoffs must thus satisfy the following individual rationality constraints for the worker (IRW) and for the firm (IRF), with $\bar{\Pi}_t$ the discounted expected future profits from a vacant job:

[4] For notational simplicity, the dependence of utility on the wage profile is not made explicit. Throughout the paper, all payoffs are functions of the agreed wage profile.

IRW:

$$U_t \geq \bar{U}_t, \text{ for all } t \geq \tau, \tag{3}$$

IRF:

$$\Pi_t \geq \bar{\Pi}_t, \text{ for all } t \geq \tau \tag{4}$$

In addition it must be in the worker's interest not to shirk and in the firm's to pay any promised bonus. For models of the present type, there typically exist many patterns of behavior that are consistent with equilibrium, including ones for which workers are not fired for shirking. It is an insight of Abreu (1988) that the set of equilibrium payoffs in a repeated game of the present type can be completely characterized in terms of the most severe penalties for cheating that are available. Since our concerns here are primarily with the set of equilibrium payoffs and the form of the wage agreement, we characterize the set of self-enforcing agreements using the most severe punishments.[5]

Given that both worker and firm can terminate the relationship at any time, the most severe penalties are given by the alternatives currently available in the market. If the reason for a separation were known to the market, it would be possible to establish a reputation with other potential partners for not cheating. We assume here that prospective employers and employees do not know why the previous matches of potential partners came to an end, so establishing an external reputation of that sort is impossible. This is necessarily the case in an anonymous market in which it is hard to keep track of participants, something that may well be true of workers from poor areas of large cities in the developing world. But even where it is easier to keep track of participants, the use of reputation depends on there being reliable sources of information about why a separation occurred — the word of mouth of parties to the match is unlikely to be reliable since, if reputation is valuable, neither has an incentive to admit to cheating. Many employers are unwilling to provide information about former employees. Bewley (1993) questioned 10 Connecticut firms about this: "nine said former employers would give them only dates of employment. They wouldn't even give salary information. Employers can be sued for giving out negative information about former employees. As a result, they don't give out positive information either, for then they can be asked another time why they are not saying anything complimentary." (p. 80). It is typically only in small, well informed markets where it is public information that a person switched jobs because of a better offer elsewhere that reputation effects become really effective.

Without external reputation effects, $\bar{U}_t$ and $\bar{\Pi}_t$ represent the current spot market alternatives no matter what the reason for separation. Thus to deter shirking by the worker, the expected future utility of keeping to the agreement, U_t, must be at least as great as the utility from shirking (with zero disutility of effort), collecting the wage w_t but not the bonus b_t, being fired, and receiving the discounted future utility $\delta\bar{U}_{t+1}$ from looking for another match at $t+1$. That is, the wage profile must generate

[5] For this model, there exist many different patterns of behavior consistent with the existence of an equilibrium, including threatening to shirk rather than threatening to quit. See MacLeod and Malcomson (1989) for a more complete discussion of the set of equilibrium agreements in a long term relationship.

payoffs that satisfy the following incentive compatibility (no shirking) condition for the worker:

$$U_t \geq w_t + \delta \bar{U}_{t+1}, \text{for all } t \geq \tau. \tag{5}$$

Substitution from (1) for the left hand side allows this condition to be rewritten

ICW:

$$E\left\{\text{future gains to worker}|t\right\} \equiv \delta\rho_{t+1}(U_{t+1} - \bar{U}_{t+1}) \geq v - b_t, \text{ for all } t \geq \tau. \tag{6}$$

The intuition for this is as follows. There are two ways to induce the worker not to shirk. One is to have a bonus b_t that compensates for the disutility of effort v. The other is to ensure that the worker has sufficient gains $\delta\rho_{t+1}(U_{t+1} - \bar{U}_{t+1})$ from continuation of employment in the future that it is worth incurring the disutility of effort in order to avoid being fired. (ρ_{t+1} is the probability that employment continues when the worker does not shirk.) If the bonus is sufficiently large then it is not necessary for the worker to earn rents from continuing the relationship. Such payments are explicitly excluded in the efficiency wage literature, which is why it is argued that workers must be offered a wage premium to discourage shirking.[6]

Since the bonus is not legally enforceable, to induce the firm to pay it the expected future profits from paying and having the match continue, Π_t, must be at least as great as from not paying (having already received the product p and being unable to escape paying the wage w_t) and having the worker quit (which results in expected future profit of $\delta\rho_{t+1}\bar{\Pi}_{t+1}$). Firm profits must therefore satisfy the incentive constraint

$$\Pi_t \geq p - w_t + \delta\rho_{t+1}\bar{\Pi}_{t+1}, \text{ for all } t \geq \tau. \tag{7}$$

Substitution from (2) for the left hand side allows this condition to be written

ICF:

$$E\left\{\text{future gains to firm}|t\right\} \equiv \delta\rho_{t+1}(\Pi_{t+1} - \bar{\Pi}_{t+1}) \geq b_t, \text{ for all } t \geq \tau. \tag{8}$$

The intuition is simply that the firm will not pay a bonus b_t unless the expected future gains from the employment exceed that bonus.

A worker's threat to quit should the firm not pay a deserved bonus may seem incredible if there is a gain from continued employment. The same applies to a firm's threat to fire a shirking worker. However, MacLeod and Malcomson (1989) show that such threats can form part of a set of self-enforcing social norms. As discussed in more detail below, what is required is for both the worker and the firm to believe that the other would continue to renege if they were to continue the relationship. In more common language, once a worker shirks or the firm refuses a deserved bonus, the relationship sours and both sides believe that it cannot continue, as discussed in the Introduction with respect to the traders at First Boston.

The expected future gains to the relationship as a whole are given by the sum of the left hand sides of ICW and ICF. When summing the right hand sides, the bonus term cancels to yield the necessary condition for a contract to be self-enforcing that

[6] See Carmichael (1990) for an extensive discussion of these issues.

IC

$$E\left\{\text{future gains to relationship}|t\right\} \equiv \delta\rho_{t+1}\left[U_{t+1}+\Pi_{t+1}-(\bar{U}_{t+1}+\bar{\Pi}_{t+1})\right] \\ \geq v, \text{ for all } t \geq \tau. \quad (9)$$

Both worker and firm are risk neutral and share the same discount rate, so these expected future gains are independent of the wage profile — the wage terms cancel when (1) and (2) are added.[7] Moreover, the necessary conditions IRW and IRF depend only on total earnings, not on how these earnings are divided between wages and bonuses. Thus, the individual rationality conditions, IRW and IRF, and the global incentive constraint, IC, are all *independent* of this division. Moreover, MacLeod and Malcomson (1989) show that these three conditions are not only necessary but also sufficient for the existence of a self-enforcing agreement. When they are satisfied, it is always possible to find a sequence of wage and bonus payments such that the worker and firm incentive constraints, ICW and ICF, are satisfied.

Since $v > 0$, an implication of the incentive constraint IC is that the existence of a self-enforcing agreement is inconsistent with both worker and firm being indifferent between the current agreement and what they might earn in the spot labor market — IC for period $t-1$ cannot be satisfied if both $U_t = \bar{U}_t$ and $\Pi_t = \bar{\Pi}_t$. Moreover, this constraint for period t depends not on the current wage and bonus payments, only on the *future* returns to the relationship. It is this fact that necessitates a social norm that may be interpreted as a fair wage. To see this, consider the position in period t of a relationship that started in period $\tau < t$. Suppose, as in the shirking model of efficiency wages, the firm receives no rent, so $\Pi_t = \bar{\Pi}_t$. By IC, the provision of incentives in period $t-1$ then requires $U_t \geq \bar{U}_t + v/\delta\rho_t$. Thus along any wage profile that provides the incentive for the worker not to shirk, the worker receives utility at the beginning of period t greater than the next best market alternative $\bar{U}_t$. However, once period $t-1$ is passed, what is important for incentives at t are any bonus payment for t and the wages and bonus payments from period $t+1$ onwards, not the wage w_t, so the firm has an incentive to negotiate w_t downwards. It could at the beginning of period t offer a new wage $w_t' < w_t$ that leaves the worker indifferent or slightly better off than at the market alternative *without affecting the incentive to perform in period t.* If, however, the worker anticipates at $t-1$ that the firm will be successful in reducing the wage at t to the market alternative, the incentives for effort at $t-1$ will be destroyed.

This problem is solved if the wage w_t has the status of a fair wage. Should the firm attempt to offer at time t a wage w_t' less than w_t, the worker responds by shirking in period t if w_t' is greater than the one period return in the spot market, or by quitting otherwise. Note that it is in the worker's interest to respond in this way if she believes that a firm that succeeds in reducing the wage in this period will also do

[7] We can see this by defining the *rent* $R_t = U_t + \Pi_t - (\bar{U}_t + \bar{\Pi}_t)$ and substituting from (1) and (2) to give

$$R_t = p - v + \delta\left(\rho_{t+1}R_{t+1} + \bar{U}_{t+1} + \rho_{t+1}\bar{\Pi}_{t+1}\right) - (\bar{U}_t + \bar{\Pi}_t).$$

Recursive substitution for future values of the rent results in the wage and bonus terms dropping out.

so in future periods. Responses of this type are in line with the view of 95% of the managers interviewed by Blinder and Choi (1990) that an unfair wage policy would result in reduced work effort. Provided the agreed wage profile yields the firm profits equal to or larger than the market alternative, the firm is hurt if the worker responds to a wage cut in this way.[8] Observe that the relationship is reciprocal in the sense that the firm offers a high wage in return for more effort. This is consistent with the experimental evidence of Fehr, Gächter and Kirchsteiger (1995) who find that the potential for reciprocal exchange increases the efficiency of contracting.

Conversely, consider the case in which the worker receives no rent, so $U_t = \bar{U}_t$. By condition IC, the provision of incentives at $t-1$ then requires that $\Pi_t \geq \bar{\Pi}_t + v/\delta\rho_t$ and the firm receives profits at t greater than the next best market alternative $\bar{\Pi}_t$. But once the firm has paid any bonus for period $t-1$, what is important for incentives are the wages and bonus payments for $t+1$ on, not the wage w_t. Thus the worker has an incentive to renegotiate the wage upwards to capture some of the firm's rent and if the firm anticipates at $t-1$ that this will be successful, the incentive to pay the bonus at $t-1$ will be destroyed. What is required in this case is a social norm ensuring that the worker cannot bid up the wage in this way.

These arguments apply as much to the beginning of the relationship as during it. If in period τ workers in the market believe that a fair starting wage is $w\tau$, even if the resulting utility is greater than the market alternative ($U\tau > \bar{U}\tau$), firms will offer $w\tau$ because they believe that a lower offer would result in shirking workers. In this way, paying an above market clearing wage can be seen as creating the expectation that the worker will reciprocate with high effort. Equally, an offer from an unemployed worker to work for less will be regarded with the suspicion that the worker intends to treat the employment as a way to make a short term gain by shirking. What is regarded as fair is a convention that coordinates behavior because it results in adverse consequences for those seen to break it, in exactly the same way as the beliefs discussed above. In all sorts of contexts, people who behave unconventionally (by, for example, offering "too good a deal") are regarded with the suspicion that "there must be a catch."

These results provide a link between the problem of worker motivation and the concept of fairness that authors such as Akerlof (1982) have argued are important in the operation of labor markets. The theory adds to the Akerlof model by introducing a set of constraints describing the set of feasible contracts. These constraints form the starting point for the discussion of market equilibrium in self-enforcing agreements in the next section.

[8] There are subtle issues to ensure that the threats that we describe are in fact credible. These issues are discussed in some detail in MacLeod and Malcomson (1989), where we show that the threats described here are credible in the sense of being part of a perfect equilibrium in a well defined employment game. See also the survey by Pearce (1992) for a review of the issues involved in sustaining cooperation.

3 Market Equilibrium

Given the spot market returns, the constraints (3), (4) and (9) completely characterize the set of wage profiles that can be supported by some self-enforcing agreement. In this section we characterize market equilibria when firms compete with each other by offering self-enforcing agreements. Though the notion of a fair wage is not explicitly discussed, the equilibrium agreements that satisfy the incentive constraints are in each case enforced by an underlying set of self-enforcing norms. Thus, while our model depends on the notion of a fair wage or bonus payment, the level of the wage or bonus is endogenously determined as a function of underlying fundamentals.

Consider first the set of equilibria in a stationary environment with a fixed number of jobs J each period, with exogenous separations at a constant rate $\rho_t = \alpha$ because it becomes unprofitable for firms to continue some jobs for reasons other than pay. The cost of creating an additional job is a constant c independent of the number of existing jobs and productivity $p(E)$ is strictly decreasing in employment E. To retain stationarity, new jobs are created at a rate equal to the rate of exit, $(1-\alpha)$ per period. In the absence of reputation effects from previous matches, in equilibrium unmatched workers and firms reach agreement immediately and thus $E = \min\{L, J\}$.

In a stationary equilibrium with firms offering the same self-enforcing agreement each period, total pay is $W = w + b$. For any W that is self-enforcing given the market alternatives $\bar{U}$ and $\bar{\Pi}$, there is always a division of pay into a base wage and bonus that satisfies the incentive constraints (6) and (8). Hence it is sufficient to describe an equilibrium state of the economy by (W, J), the total pay for each job and the number of jobs. Let $U(W,J)$ and $\bar{U}(W,J)$ denote the expected future utility in equilibrium of an employed worker and of a worker entering the labor market, respectively, and $\Pi(W,J)$ and $\bar{\Pi}(W,J)$ the equilibrium expected future profits from a filled job and from a vacant job, respectively. For a stationary equilibrium, it follows from (1) and (2) that worker utility and firm profits from employment under the present assumptions are

$$U(W,J) = W - v + \delta\left[\alpha U(W,J) + (1-\alpha)\bar{U}(W,J)\right], \quad (10)$$

$$\Pi(W,J) = \left[p(E) - W\right]/(1-\delta\alpha). \quad (11)$$

Recall that $\Pi(W,J)$ is the *ex post* profit from a job after the cost of creating it has been incurred. With separations occurring because jobs become unprofitable with probability $1-\alpha$, the effective discount factor for firms is $\delta\alpha$.

There can be no equilibrium at which the number of jobs J is equal to the number of workers L. With equal numbers, all workers and jobs are matched every period, so $U(W,J) = \bar{U}(W,J)$ and $\Pi(W,J) = \bar{\Pi}(W,J)$. The rent from continuing a match is then zero, so it is not possible to satisfy the incentive compatibility constraint (9) and no self-enforcing agreement exists. As in Shapiro and Stiglitz (1984), the rent needed for a set of self-enforcing social norms is generated by an imbalance between demand and supply in the labor market. However, as MacLeod and Malcomson (1989) observe, this rent can be generated by unfilled vacancies just as well as by an excess supply of workers. We therefore consider the two possibilities $J < L$ and $J > L$. When $J < L$ we show that an efficient equilibrium involves the use of efficiency wages: workers are paid an above market clearing wage and fired if they shirk. Conversely, when $J > L$, efficient equilibria involve the use of bonus payments to workers who do not shirk.

3.1 Efficiency Wage Equilibria

Consider first the possibility of equilibrium with the number of jobs J less than the number of workers L. In this case, $E = J$ and vacant jobs can always be filled straightaway, so they have the same expected future profits as filled jobs. Thus

$$\bar{\Pi}(W, J) = \Pi(W, J) = \left[p\left(J\right) - W\right]/(1 - \delta\alpha), \tag{12}$$

where the second equality follows from (11). To replace jobs that become unprofitable $(1 - \alpha)J$ vacancies are created each period. Thus the probability of an unemployed worker finding a job in any period is $\pi(J) = (1 - \alpha)J/(L - \alpha J)$. For given (W, J), the equilibrium utilities of employed and unemployed workers are then given by the unique solution to (10) and

$$\bar{U}\left(W, J\right) = \pi\left(J\right)U\left(W, J\right) + \left[1 - \pi\left(J\right)\right]\left[u + \delta\bar{U}\left(W, J\right)\right], \tag{13}$$

where u is the utility during one period of unemployment. The utility of both employed and unemployed workers is increasing in W and J.

Two conditions must be met for (W, J) to be a stationary equilibrium. First, firms must be prepared to create enough new jobs to replace jobs that become unprofitable for exogenous reasons but no more, which requires that $\Pi(W, J)$ equals the cost c of creating an additional job. The equilibrium entry condition is thus

$$\left[p\left(J\right) - W\right]/(1 - \delta\alpha) = c. \tag{14}$$

Second, neither firms nor workers must be able to gain by cheating on their agreement. Since firms can always fill vacancies, they would always cheat on a bonus. (Formally, with $\Pi(W, J) = \bar{\Pi}(W, J)$, (8) requires that $b \leq 0$ but nothing is lost by setting the bonus to zero.) Thus bonus payments cannot be part of an equilibrium and the only pay is the base wage w. With $b = 0$, the worker's incentive compatibility constraint (6) reduces to the constraint

$$U(W, J) - \bar{U}(W, J) \geq v/\delta\alpha. \tag{15}$$

Since $v > 0$, this constraint implies that employed workers have strictly higher expected future utility than unemployed workers, so the base wage w must be above the market clearing level. That is, it is an efficiency wage. Substitution of the solutions for $U(W, J)$ and $\bar{U}(W, J)$ derived from (10) and (13) and some tedious algebra allows (15) to be rewritten as the *no shirking constraint* for workers

NSC:

$$W \geq u + \frac{v}{\left[1 - \pi\left(J\right)\right]\delta\alpha}. \tag{16}$$

A market equilibrium is a pair (W, J) that satisfies the entry and no shirking constraints (14) and (16). These constraints are illustrated in figure 2. The entry condition (14) implies that an equilibrium must lie on the line $W = p(J) - (1 - \delta\alpha) c$ labelled *zero profits*, which is downward sloping because $p(J)$ is decreasing in J. The no shirking constraint (16) implies that an equilibrium point must lie above the upward sloping line labelled NSC that starts from $u + v/(\delta\alpha)$ for $J = 0$ and is asymptotic to the line $J = L$. (If there are no jobs, the probability $\pi(J)$ of a worker finding a job is zero and, as the number of jobs approaches the number of workers, $\pi(J)$ approaches 1. This line is also everywhere upward sloping and strictly convex.) Thus the set of efficiency wage equilibria corresponds to the heavy part of the zero profits line to the left of the point A in figure 2.

Despite free entry, the potential for the creation of self-enforcing social norms implies that individual incentives are not sufficient to determine a unique equilibrium. Each point in the set of efficiency wage equilibria corresponds to a different fair wage convention. For example, point E in figure 2 corresponds to all firms offering the wage w_E. Should a firm attempt to enter with a lower wage, workers would interpret such a wage as unfair, and shirk if employed at that wage. Thus sustaining efficiency wage equilibria requires market conventions like those found by Blinder and Choi (1990) that prevent firms cutting wages simply because there is excess supply of labor.

Equilibria with more jobs have higher employment and so exploit more gains from trade. Following Hayek (1982), suppose that competition, experimentation and evolution result in the market discovering and adopting a set of conventions yielding an outcome that is efficient in the sense that all potential gains from trade are exploited.[9] At the moment there is no generally accepted approach to modelling such a complex process and we do not attempt it here. Rather we begin with the properties of an efficient equilibrium and ask which set of conventions supports such an equilibrium. Methodologically, this is not fundamentally different from the standard practice of analyzing the properties of a competitive equilibrium despite not having modelled the process by which equilibrium is reached.

For efficiency wage equilibria, efficiency corresponds to maximizing employment. Thus the efficient equilibrium is at point A in figure 2, the intersection of the no shirking constraint and the zero profits line, with employment J_A. This is precisely the equilibrium identified by Shapiro and Stiglitz (1984) for the case $c = 0$. In that case, the zero profits line can be identified with aggregate labor demand.[10]

How does such a market respond to shocks? Suppose there is an unanticipated but permanent shock that reduces $p(J)$. This shifts the zero profit line downwards. Exactly as in Shapiro and Stiglitz (1984), the no shirking constraint is independent of

[9] Essentially this is a notion of *inclusive fitness:* the institutions we observe survive because they help the group survive against competitors. See Boyd and Richerson (1985) for further discussion and additional references.

[10] Since firms receive zero profits in every equilibrium in figure 2 and purchasers of the product receive more surplus from higher output, the equilibrium at A will Pareto dominate if workers have higher utility. Workers who would be employed in an equilibrium with lower employment may lose out because the wage is lower, although this is at least partially offset by a higher probability of re-employment when their current job ends. Whether workers on average (or behind the veil of ignorance) gain depends on how steeply $p(J)$ declines as J increases.

$p(J)$ for given J, so the shock shifts point A downwards along the NSC curve. Thus the new long run equilibrium has lower wages and employment even though labor supply is completely inelastic. But, if $c > 0$, the move to the long run equilibrium can be delayed. Because the cost of creating the J_A jobs has already been incurred, firms will continue to fill them as long as $p(J_A) \geq W$, though they will not create replacement jobs as current jobs disappear. Indeed, for a shock that reduces $p(J_A)$ by less than $(1-\delta\alpha)\, c$, employment and wages need not adjust at all initially. The number of jobs will decline as those that become unprofitable are not replaced until the new long run equilibrium number is reached. At that point wages must fall to induce firms to create replacement jobs. How wages evolve in the meantime depends on how market conventions evolve in response to shocks.

3.2 Performance Pay Equilibria

Consider now the possibility of equilibria with more jobs than workers $(J > L)$ so that $E = L$. In this case, workers without a job can always get one straightaway and so have the same expected future utility as those with a job. Thus

$$\bar{U}(W, J) = U(W, J) = (W - v)/(1 - \delta), \tag{17}$$

the second equality following from use of the first in (10). Employed workers then have the same expected utility as unemployed workers, so efficiency wages are not paid. Exogenous turnover generates $(1-\alpha)\, L$ workers seeking new jobs each period. Thus the probability of a vacancy being filled in any period is $\gamma(J) = (1-\alpha)L/\left(J - \alpha L\right)$. For given (W, J), the equilibrium *ex post* profits from filled and from vacant jobs are given by the unique solution to (11) with $E = L$ and

$$\bar{\Pi}(W, J) = \gamma\left(J\right)\Pi(W, J) + \left[1 - \gamma(J)\right]\delta\alpha\bar{\Pi}(W, J), \tag{18}$$

which gives the expected future profits from a vacancy given the probability $\gamma(J)$ of filling it in each period and the discount factor for firms $\delta\alpha$.

As with efficiency wage equilibria, for (W, J) to be an equilibrium firms must not want to change the number of jobs. Thus the post entry profit per worker must equal the cost of creating a new job:

$$\bar{\Pi}(W, J) = c. \tag{19}$$

Moreover, neither firms nor workers must gain by cheating on their agreement. Since workers can always get another job, they will shirk unless there is a bonus of amount v to compensate for the disutility of effort, so any equilibrium of this type must have performance pay. (Formally, with $U(W, J) = \bar{U}(W, J)$, (6) requires $b \geq v$ but nothing is lost by setting $b = v$.) Substitution of this into the firm's incentive compatibility condition ICF (8) gives the *no cheating constraint* for firms

NCC:

$$\Pi(W, J) - \bar{\Pi}(W, J) \geq v/\delta\alpha. \tag{20}$$

Constraints (19) and (20), along with the corresponding efficiency wage constraints for comparison, are illustrated in figure 3. The left half of the figure reproduces the curves from figure 2. The zero profit line continues to fall as the number of jobs is increased beyond the number of workers L due to an increase in vacancies and thus a reduction in the probability of finding a worker for the job. The no cheating constraint (20) implies that an equilibrium must lie below the upward sloping line labelled NCC. This line slopes upward because, with more jobs, it takes longer for a firm to fill a vacancy, thus reducing the profits required to prevent reneging on the bonus $b = v$ and increasing what can be paid to workers. It cannot cross the line $J = L$ because with $J = L$ a firm could always replace a worker immediately and would therefore never pay a bonus. The set of performance pay equilibria corresponds to the heavy part of the zero profits line (to the right of point B).[11] In any such equilibrium, firms with jobs that have already been created receive higher profit than they require to continue those jobs because $W < p(L)$. To sustain this requires a convention preventing workers bidding up wages even though there are unfilled vacancies that is exactly symmetric to the convention that prevents firms bidding down wages in efficiency wage equilibria even though there are unemployed workers.

As in the efficiency wage case, individual incentives are not by themselves sufficient to determine a unique equilibrium. All performance pay equilibria have the same employment L and the same output. Due to the cost of replacing jobs that become unprofitable for reasons other than pay, an efficient equilibrium minimizes the number of unfilled vacancies. Thus the unique efficient performance pay equilibrium is that given by B in figure 3 with J_B jobs. It Pareto dominates other performance pay equilibria since output and employment are the same, wages are higher, and firms receive zero profits in all such equilibria.

Productivity shocks affect the efficient performance pay equilibrium in a particularly simple way. It can be shown that a reduction in $p(L)$ reduces both the NCC and the zero profits lines by the same amount for given J. In consequence an unanticipated but permanent fall in $p(L)$ leaves the efficient number of jobs unchanged but requires a fall in total pay. Thus pay responds to such a shock but employment and vacancies remain unchanged unless the shock is sufficiently adverse that a performance pay equilibrium no longer exists. If it is sufficiently adverse, there must be a transition to an efficiency wage equilibrium with a consequently large fall in employment but the dynamics of how that happens depend on how market conventions evolve in response to shocks.

3.3 Efficiency Wage versus Performance Pay Equilibria

One can compare the performance pay equilibrium with J_B jobs to the efficiency wage equilibrium with employment J_A. The cost of moving from the efficiency wage equilibrium with J_A jobs to the performance pay equilibrium with J_B jobs and employment L is the cost of creating the additional jobs which, amortized on a per period basis, is $(1-\delta\alpha)(J_B - J_A)c$, plus the additional disutility of work $(u+v)(L-J_A)$. The benefit from the move is the benefit from the additional output from increasing

[11] Performance pay equilibria exist if $p(L) - (1-\delta\alpha)c \geq u + v/\delta\alpha$.

employment from J_A to L. Thus there is a net gain from the move if and only if

$$\int_{J_A}^{L} p(J)dJ > (1-\delta\alpha)(J_B - J_A)c + (u+v)(L - J_A). \tag{21}$$

When it is costless to create a job ($c = 0$) as in Shapiro and Stiglitz (1984), the performance pay equilibrium is more efficient whenever the value of employment $p(L) - v$ is greater than the gain from unemployment u, the same condition (marginal product greater than disutility of work) for it to be efficient to employ all workers even if there were no problem of contracting on effort. If either the cost of creating a job is sufficiently high, or the marginal product $p(L)$ is sufficiently elastic, then the efficiency wage equilibrium is more efficient.

One result from the analysis is that, while efficiency wage equilibria have unemployment, performance pay equilibria do not. This points to a potentially testable prediction. Bonus payments in the model are different from other forms of incentive pay such as piece rates in being in effect discretionary, not contractually fixed. Previous work, such as Brown (1990), has focused on pay systems that link pay formally to some performance measure. Yet in many occupations it is common practice to make performance pay discretionary, as with the traders at First Boston discussed in the Introduction. The model predicts that if the unemployment rate for an occupation is low, we should observe the use of such bonus payments, while in a high unemployment environment we should observe efficiency wages.

Unfortunately standard data sets such as the Panel Study of Income Dynamics (PSID) or the National Longitudinal Survey of Youth (NLSY) do not distinguish between discretionary payments and contractually set incentive pay. One data set that does is the 1990 British Workplace Industrial Relations Survey. Though it does not report the size of discretionary merit pay, it does explicitly ask if a worker received merit pay, as opposed to performance pay with "a mechanical relationship between the worker's pay and some relatively objective measure of output" (Millward et al. (1992, p. 258)). The above analysis suggests that, in data grouping workers in different labor markets, the incidence of merit pay should be higher for groups with a lower unemployment rate. In figure 4 we plot the incidence of merit pay by broad occupation against the occupational unemployment rate taken from the 1990 British Labour Force Survey (with the line representing the OLS fit that has slope -2.81 and t-ratio -15.95). The relationship is negative, as predicted. Note that the incidence of merit pay varied from 10% to 35%. Of course this is not a test of the theory. That would require better data and a carefully specified econometric model. It does, however, illustrate that merit pay is a phenomena that is empirically relevant, as well as one about which theory can make predictions.

The property that equilibria in the model have all rent going to the long side of the market (to workers if there is unemployment, to firms if there are unfilled vacancies) applies to new matches even if they are not required to offer the same agreements as current matches and also to non-stationary environments. The essential intuition is that a firm or an employee receiving no rent from a current match cannot receive a rent from a new match without making it profitable to renege on its current agreement. The details are in an appendix.

4 Equilibrium in a Two Sector Economy

4.1 Capital Intensity and Wages

This section illustrates in a simple two sector economy how the relative capital intensity of jobs has implications both for the level of wages and for the relationship between wages and turnover. A number of studies, including Krueger and Summers (1988), Abowd, Kramarz and Margolis (1993) and Card and Krueger (1995) suggest that there are variations in wages across markets and industries that cannot be explained by the standard competitive models. Dickens and Katz (1987) find that high wages are associated with more profitable capital intensive industries. A simple rent sharing argument would suggest that *all* workers in such firms earn higher wages. However, recent work by Davis and Haltiwanger (1991) using matched worker-firm data finds that production workers, rather than non-production workers, account for much of the inter-industry wage variation. Troske (1994) finds that much of this inter-industry wage variation for production workers can be explained by the capital intensity of the job.[12]

We capture the distinction between relatively capital intensive and relatively labor intensive jobs in its starkest form with the following assumptions. Capital intensive jobs are characterized by a limited number J_t of positions in period t that can be increased only at a sunk cost of $c_t(J_t)$ per job as discussed in the previous section. We assume $c_t(J_t)$ increases sufficiently with J_t as a result of diminishing returns, congestion costs, etc., that the number of production jobs is bounded and less than the number of available workers. Each capital intensive job has productivity (net of non-labor costs) of p^k independent of the number of jobs filled. In contrast, there are no sunk cost to creating additional labor intensive jobs so, even in the short run, there is a potentially unlimited number of such jobs. Assuming an unlimited number simplifies the analysis without affecting its qualitative features — the essential characteristic is that workers are on the short side of the market in this sector. Each labor intensive job has productivity (net of non-labor costs) of $p^\ell < p^k$. To highlight the effect of capital intensity, it is assumed that both sectors face exactly the same problem of worker motivation.

Workers are freely mobile between the two types of jobs, so the probability ρ_t^i that a job of type i continues in period t must now reflect not only exogenous separations but also quits to the other type of job. Free mobility implies that unemployed workers have the same utility no matter what type of job they have been searching for, so $\bar{U}_t^k = \bar{U}_t^\ell$.

Under these conditions, an efficient equilibrium has all workers in the labor intensive sector employed in every period. Thus these workers can always find a labor intensive job, implying $\bar{U}_t^\ell = U_t^\ell$. Together with free mobility, this implies

$$U_t^\ell = \bar{U}_t^\ell = \bar{U}_t^k. \tag{22}$$

The incentive condition IC (9) then implies that labor intensive jobs with a worker

[12] See Kremer (1993) and Kremer and Maskin (1994) for an alternative competitive explanation for these observations.

must yield a rent to firms in period t.[13] With an unlimited number of such jobs, the expected profits from jobs without a worker are zero ($\bar{\Pi}_t^\ell = 0$). Together these imply $\Pi_t^\ell = v/\delta\rho_t^\ell$. If the market were perfectly competitive, workers would receive all this rent and hence equilibrium pay would be $W^* = p^\ell$. Use of these in the expression for profits (2) in section 2 implies that equilibrium pay in labor intensive jobs is

$$W_t^\ell = W^* - v\left[(1-\delta\rho_t^\ell)/\delta\rho_t^\ell\right]. \tag{23}$$

Workers give up a rent of $v[(1-\delta\rho_t^\ell)/\delta\rho_t^\ell]$ each period that ensures the firm has an incentive to pay the end of period bonus should the worker perform well.

Because capital intensive jobs have higher productivity than labor intensive jobs, it is efficient to have all these jobs filled every period. In equilibrium, capital intensive jobs can thus fill vacancies immediately and so, in contrast to labor intensive jobs, firms do not face a loss in profits when a worker leaves. Therefore $\Pi_t^k = \bar{\Pi}_t^k$.[14] In this case to ensure that the incentive constraint IC (9) is satisfied, employed workers must earn a rent $v/\delta\rho_t^k$ from employment and hence

$$\begin{aligned} U_t^k &= \bar{U}_t^k + v/\delta\rho_t^k, \qquad (24) \\ &= U_t^\ell + v/\delta\rho_t^k. \end{aligned}$$

Here the second line follows from (22).

Expressions (23) for pay in labor intensive jobs and (24) for utilities in the two sectors can be used in the equation for worker utility (1) from section 2, together with (22), to show that pay in capital intensive jobs is

$$\begin{aligned} W_t^k &= W_t^\ell + v\left(\frac{1-\delta\rho_t^k}{\delta\rho_t^k}\right), \qquad (25) \\ &= W^* + v\left[\left(\frac{1-\delta\rho_t^k}{\delta\rho_t^k}\right) - \left(\frac{1-\delta\rho_t^\ell}{\delta\rho_t^\ell}\right)\right]. \qquad (26) \end{aligned}$$

This expression relates the size of the wage differential to the level of turnover on the job and the disutility of effort. As the probability of turnover decreases for capital intensive jobs, so does the wage differential ($\partial(W_t^k - W_t^\ell)/\partial\rho_t^k < 0$). The wage differential does not depend on the turnover level for labor intensive jobs. In the case of labor intensive jobs, turnover has the opposite effect, with wages increasing as turnover falls ($\partial W_t^\ell/\partial\rho_t^\ell > 0$). In contrast to the theory of compensating differentials, this model illustrates that there may be no consistent relationship between turnover and wages. The theory does imply, however, that in sectors of the economy where bonus payments are observed there should be a negative relationship between turnover and wages, and conversely when workers are paid a fixed wage.

The essential characteristics of capital intensive jobs in the model are that (i) they have higher productivity, and (ii) there is a sunk cost to creating vacancies that

[13] When turnover occurs because workers quit, not just because jobs become unprofitable, equations (2) and (7) in section 2 need some amendments when $\bar{\Pi}_{t+1} \neq 0$, but these amendments cancel out in deriving (8) and (9).

[14] In practice there are always some costs to filling a vacancy, such as training. In this section we are concerned with search costs arising from the time it takes to find a suitable replacement.

cannot be recouped if the vacancy is not filled. This fits better with the descriptions of Ford's $5 day example in Raff and Summers (1987) and Raff (1988) than does the shirking story in Bulow and Summers (1986) in which efficiency wages are higher where monitoring is more difficult. Raff (1988) comments that the shift to assembly line work that Ford had adopted must have made at least some aspects of monitoring easier (for example, speed of work) and that there was actually an increase in the ratio of supervisory to production employees in the aftermath of the introduction of the $5 day. But Raff (1988, p. 395) notes that "the technical change that made monitoring easier also involved extraordinarily highly dedicated physical capital and a production process that sent unprecedented numbers of pieces past each worker in each unit of time." This made the cost to Ford of having a work station not effectively operated high. It is precisely this cost of not having each position filled all the time that makes it efficient in the present model for the market to allocate the rent required for a self-enforcing agreement to employees — for jobs to have a rent requires them to remain unfilled for a time whenever a match ends.

The analysis applies to both the case in which workers can search for higher paying capital intensive jobs while employed in labor intensive jobs and the case in which they must quit their labor intensive job before they can search for a capital intensive one. There are two main differences between these cases: (i) the amount of unemployment; and (ii) the labor turnover rate in labor intensive jobs. When workers in labor intensive jobs can search while employed, there is no unemployment because, if unsuccessful in getting a capital intensive job in one period, they can always continue in their labor intensive job without prejudicing their chance of getting a capital intensive job in the next. When they cannot search on the job, those who are unsuccessful in getting a capital intensive job remain unemployed for that period. In that case, the model generates unemployment in the capital intensive sector, just as in the urban sector of the Harris-Todaro model, despite wages being determined endogenously rather than specified exogenously. Moreover, unlike in the reformulation of the Harris and Todaro (1970) model due to Moene (1988), the wage differential does not depend on differences between the two sectors in the ability to monitor performance, only on differences in capital intensity.

4.2 Equilibrium and Growth

In this section we simulate some of the dynamic properties of the model for the Harris-Todaro case in which workers must quit jobs in one sector to look for jobs in the other. In this case labor intensive jobs may be interpreted as employment in rural areas, either in agriculture or in a small enterprise involving a minimal amount of capital investment. Capital intensive jobs may be interpreted as the urban manufacturing sector.

Because productivity in capital intensive jobs is higher, all those jobs are filled in an efficient equilibrium and, as discussed in section 4.1, they have higher pay than labor intensive jobs. With search on the job not possible, workers enter the capital intensive sector by moving into the pool of the unemployed. In our simulation we consider the effect of increasing the number of capital intensive jobs and the impact that the resulting migration from the labor intensive sector has on output, unemployment, wages and inequality.[15]

In both sectors some labor turnover occurs because jobs become unprofitable for exogenous reasons. The probability of a match being terminated in each period for such reasons is denoted $(1-\alpha^i) \in [0,1)$, for $i \in \{k, \ell\}$. Given that the number of capital intensive jobs is increasing and that they have higher pay, there is no reason for endogenous turnover in this sector, so $\rho_t^k = \alpha^k$ for all t. Migration is independent of exogenous separation, so the probability ρ_{t+1}^ℓ that a labor intensive match that exists in period t continues in period $t+1$ is

$$\rho_{t+1}^\ell \equiv \alpha^\ell \cdot L_{t+1}^\ell / L_t^\ell, \tag{27}$$

where L_t^ℓ is the number of workers in the labor intensive sector (which is non-increasing over time due to the growth in the number of capital intensive jobs).

The time path of workers in the labor intensive sector, $\hat{L}^\ell \equiv \{L_t^\ell\}_{t=0}^\infty$, determines the rate of turnover from (27) and hence the wage rate each period from expression (23). Because moving to the capital intensive sector requires joining the pool of unemployed with the same expected utility in equilibrium as staying in the labor intensive sector, the equilibrium utility in the labor intensive sector as a function of the time path of workers in that sector is

$$\begin{aligned} \bar{U}_t^\ell(\hat{L}^\ell) &= U_t^\ell(\hat{L}^\ell) = \sum_{j=t}^\infty (W_j^\ell - v)\delta^{j-t} \qquad (28) \\ &= \sum_{j=t}^\infty \left(p^\ell - \frac{vL_{j-1}^\ell}{\delta \alpha^\ell L_j^\ell} \right) \delta^{j-t}, \qquad (29) \end{aligned}$$

the first equality following from (22), the last from (23).

In the capital intensive sector, there are $L_t^k \equiv L - L_t^\ell$ workers at the beginning of period t and $\alpha^k J_{t-1}$ jobs remaining filled from the previous period. Thus the probability π_t that an unemployed worker gets a job in period t is given by the number of vacancies $(J_t - \alpha^k J_{t-1})$ divided by the number of unemployed job seekers $(L_t^k - \alpha^k J_{t-1})$. Hence,

$$\pi_t = \frac{J_t - \alpha^k J_{t-1}}{(L - L_t^\ell) - \alpha^k J_{t-1}}. \tag{30}$$

The default utility for workers entering the queue for capital intensive jobs is then

$$\bar{U}_t^k(\hat{L}^\ell) = \pi_t U_t^k(\hat{L}^\ell) + (1-\pi_t)\left[u^k + \delta \bar{U}_{t+1}^k(\hat{L}^\ell)\right], \tag{31}$$

where u^k is the utility received from one period of unemployment in this sector. Together (24), (29), (30) and (31) determine the expected lifetime utility of employed and unemployed workers as functions of the numbers of workers in the labor intensive sector and jobs in the capital intensive sector in each period. The time path for workers in the labor intensive sector as a function of the number of capital intensive

[15] Formally, because the number of capital intensive jobs is endogenous to the model, the increase should be seen as the result of a fall in the cost of creating those jobs.

jobs can then be solved using the equilibrium condition that workers migrate until they are indifferent between sectors

$$\bar{U}_t^k(\hat{L}^\ell) = \bar{U}_t^\ell(\hat{L}^\ell), \text{ for all } t. \tag{32}$$

The effects of increasing the number of capital intensive jobs are simulated for rapid growth that takes the capital intensive sector from 20% of the work force to about 80% over a period of slightly less than 10 years. The productivity of capital intensive jobs is assumed to be 50% higher than that of labor intensive jobs, so the productive potential of the economy grows about 25% in the ten years. The parameter values used in the simulation are: $v = 1$, $\delta = 0.91$, $\alpha^k = \alpha^\ell = 0.9$, $p^k = 4.5$, $p^\ell = 3$, $u^k = 0$.[16] Figure 5 depicts the effect of this growth on unemployment, output (GNP), wages, and inequality of worker incomes as measured by the Gini coefficient. The dashed line represents the number of capital intensive jobs at each date. In the short run the unemployment rate rises dramatically from about 7% to almost 25% at its peak, while wages in capital intensive jobs fall by about 20%. After rising slowly, GNP dips slightly and there is a marked increase in inequality. But then GNP starts to rise sharply and, in the long run, wages and the unemployment rate return to their pre-growth levels. The effect of growth is to slightly raise inequality in the long run.

In the simulations, increased turnover in labor intensive jobs as workers move to the capital intensive sector lowers the value of a match to firms with labor intensive jobs. Consequently, pay must fall to ensure sufficient rent to firms in that sector for the wage agreement to be self-enforcing. This reduces utility in labor intensive jobs. In equilibrium, the expected utility from search in the two markets must be equal, so growth also decreases pay and utility in capital intensive jobs in the short run. This can be achieved only by a short term reduction in the probability of an unemployed worker getting a capital intensive job. Note that we do not *assume* workers *have* to spend some time unemployed after migrating to the capital intensive sector before they can get a capital intensive job. Lucky migrants get jobs without being unemployed at all. The increased unemployment is simply what is required to keep the efficient equilibrium payoff conditions satisfied.

Once growth in the number of capital intensive jobs begins to slow, migration falls and wages start to increase. Once growth stops, the unemployment rate in the capital intensive sector and real pay in both sectors returns to its original value. Growth cannot increase pay in either sector as long as there are workers in labor intensive jobs to be drawn into the capital intensive sector since the labor intensive sector has constant returns to scale. However, those migrants who succeed in getting capital intensive jobs earn higher wages than workers who remain in the labor intensive sector, so average wages increase and there is an increase in inequality.

If growth in the number of capital intensive jobs continued indefinitely, there would eventually be no workers employed in the labor intensive sector, that sector would disappear, and wages in capital intensive jobs would start to rise. Wages could rise before this if technological progress increased productivity in the labor intensive sector because by (23) W_t^ℓ increases by the same amount as p^ℓ and that feeds through

[16] The simulation was programmed in matlab. The code used is available upon request.

to wages in capital intensive jobs. This is a source of real wage growth we have ignored in the simulation that could ameliorate, and possibly outweigh, the absolute decline in real wages resulting from increased migration.

As in Harris and Todaro (1970), migration and urban unemployment are the consequences of growth in the limited number of jobs in the high wage manufacturing (capital intensive) sector in urban areas. The important difference is that here we have modelled the reasons for the wage differential between urban and rural areas and this allows us to model the consequences of urban growth for wages. Such growth, as we have shown, results in a fall in real wages in both sectors and so, at least in its early stages, results in a lowering of living standards for all workers except those who migrate and are successful in getting one of the increased number of urban jobs. In consequence, the share of profits increases and, if profits go to the already better off, growth generates the Kuznets (1955) effect of an increase in the inequality of the income distribution.

5 Conclusion

Shirking models of efficiency wages suppose that firms motivate workers by paying above market clearing wages and threatening to fire workers who shirk. Such models explain neither the widespread use of performance related pay nor the social norms surrounding the concept of a fair wage. We have shown that the theory of self-enforcing agreements can extend the basic efficiency wage model to incorporate both these phenomena.

Our model predicts that efficiency wages are likely to be observed in industries where the cost of having a job vacant is high relative to the cost of a worker remaining unemployed (or underemployed). We show that the existence of an efficiency wage equilibrium depends on a set of self-enforcing social norms that ensure firms do not cut wages purely to take advantage of an excess supply of workers, though wages still fall in response to an *increase in* excess supply. Where there is excess demand for workers, the model predicts that firms use some form of performance pay to motivate employees. In this case, the threat of a worker quitting ensures that firms have an incentive to pay the promised bonuses.

These results are applied to a two sector model and additional empirical implications are derived. It is shown that inter-sectoral wage differentials arise from differences in the capital intensity of jobs. We illustrate how a period of rapid growth in capital intensive jobs may generate an initial fall in output, reduced wages, higher unemployment and increased inequality, results that do not follow from a standard textbook model.

The model presented here is very stylized. It does, however, illustrate that viewing the problem of worker motivation from the more general perspective of self-enforcing agreements generates implications that are different from other incentive models. There are many avenues for extensions of this work, including incorporating variations in worker ability and search costs, that should result in new and interesting testable hypotheses.

A Non-Stationary Equilibria

This appendix shows that the property that equilibria in the model have all the rent going to the long side of the market (to workers if there is unemployment, to firms if there are unfilled vacancies) applies to all equilibria, not just the stationary equilibria discussed in section 3. The absence of reputation effects from previous jobs ensures that all agents on the short side negotiating new matches in period t receive the same payoff — because they can always find another match, they will refuse any offer that gives a payoff lower than the best currently available. Thus, if there are more workers than jobs, firms can always fill vacant jobs, so their future profits if they end a match are just the future profits from forming a new one. Formally, if $L > J_t$, then $\bar{\Pi}_t = \Pi_t(\hat{W}_t)$, where $\hat{W}_t$ is the pay profile negotiated for matches started at t. Conversely, if there are more jobs than workers, workers can always find another job so their future utility from ending a match is just the future utility they would get from forming a new one. That is, if $L < J_t$, then $\bar{U}_t = U_t(\hat{W}_t)$. These properties follow simply from the absence of external reputation effects and all short side agents being matched.

To satisfy the incentive compatibility (IC) condition (9) requires a strictly positive rent of at least $v/(\delta\rho_t)$ for matches formed *before* t. (At $t = 1$, the first period the market operates, there are no existing matches for which incentive compatibility must be satisfied.) That, in turn, implies the same rent for new matches started at $t > 1$. This can be seen as follows. In adding (1) and (2) for any τ, the pay terms cancel, which implies $U_t(\hat{W}\tau) + \Pi_t(\hat{W}\tau)$ is independent not only of the wage profile but also of the date τ at which a match was formed. Moreover, since the absence of external reputation effects implies that the market alternatives $\bar{U}_t$ and $\bar{\Pi}_t$ at each t are the same for all matches no matter when they were formed, it follows that

$$U_t(\hat{W}_t) + \Pi_t(\hat{W}_t) = U_t(\hat{W}\tau) + \Pi_t(\hat{W}\tau),\ \text{for all}\ \tau \leq t. \tag{33}$$

Together with the incentive compatibility condition (9), this implies

$$U_t(\hat{W}_t) + \Pi_t(\hat{W}_t) - \bar{U}_t - \bar{\Pi}_t \geq \frac{v}{\delta\rho_t},\ \text{for}\ t > 1, \tag{34}$$

which establishes that there is a rent of $v/\delta\rho_t$ to forming new matches at $t > 1$.

Consider a market equilibrium with $J_t < L$. As already shown, this has $\bar{\Pi}_t = \Pi_t(\hat{W}_t)$. But then (34) implies $U_t(\hat{W}_t) \geq \bar{U}_t + \frac{v}{\delta\rho_t}$ for $t > 1$. Thus workers necessarily receive higher utility from getting a job than from being unemployed even at the start of a match. Conversely, a market equilibrium with $J_t > L$ has $\bar{U}_t = U_t(\hat{W}_t)$ and use of this in (34) establishes that $\Pi_t(\hat{W}_t) \geq \bar{\Pi}_t + \frac{v}{\delta\rho_t}$ for $t > 1$. Thus firms necessarily receive higher profits from filling a job than from leaving it vacant even at the start of a match. To summarize, a market equilibrium with more jobs than workers has all workers employed and firms receiving a rent of at least $v/(\delta\rho_t)$ from filling a vacant job at any $t > 1$. In contrast, a market equilibrium with more workers than jobs has all jobs filled and workers receiving a rent of at least $v/(\delta\rho_t)$ from becoming employed at any $t > 1$.

This argument applies only to matches formed in periods $t > 1$, that is, to periods after the first period in which the market operates. It does not apply to matches formed in period 1 because decisions to shirk and cheat depend only on future payoffs and so are unaffected by payoffs in period 1. (This is just the point made in Carmichael (1985) about bonding.) But, in any on-going market, $t = 1$ is a date in the past and, as time goes by, matches formed at that date become increasingly irrelevant to current market conditions because of exogenous turnover.

References

Abowd, John, Francis Kramarz, and David Margolis, "High Wage Workers and High Wage Firms," 1993. Cornell University.

Abreu, Dilip, "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica*, March 1988, *56* (2), 383–396.

Akerlof, George A., "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 1982, *97*, 543–569.

Bewley, Truman F., "A Depressed Labor Market, as Explained by Participants," February 1993. Unpublished, Yale University.

Blinder, Alan S. and Don H. Choi, "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics*, November 1990, *105* (4), 1003–1015.

Boyd, Robert and Peter Richerson, *Culture and the Evolutionary Process*, Chicago: University of Chicago Press, 1985.

Brown, Charles, "Firms' Choice of Method of Pay," *Industrial and Labor Relations Review*, February 1990, *43* (3 (Special Issue)), 165S–182S.

Bulow, Jeremy I. and Lawrence H. Summers, "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment," *Journal of Labor Economics*, July 1986, *4* (3, pt 1), 376–414.

Card, David and Alan B. Krueger, *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton, NJ: Princeton University Press, 1995.

Carmichael, H. Lorne, "Can Unemployment Be Involuntary?: Comment," *American Economic Review*, December 1985, *75* (5), 1213–1214.

____ , "Efficiency Wage Models of Unemployment: One View," *Economic Inquiry*, April 1990, *28*, 269–295.

Davis, Steven J. and John Haltiwanger, "Wage Dispersion Between and Within U.S. Manufacturing Plants, 1963–1986," Technical Report, NBER 1991.

Dickens, William and Lawrence F. Katz, "Inter-Industry Wage Differentials and Industry Characteristics," in Kevin Lang and Jonathan Leonard, eds., *Unemployment and the Structure of Labor Markets*, London: Basil Blackwell, 1987.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger, "Reciprocity as a Contract Enforcement Device: Experimental Evidence," August 1995. University of Zürich.

Harris, John R. and Michael P. Todaro, "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review*, March 1970, *60* (1), 126–142.

Hayek, F. A., *Law, Legislation and Liberty*, London: Routledge and Kegan Paul, 1982.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, *76* (4), 728–741.

Kaufman, Roger T., "On Wage Stickiness in Britain's Competitive Sector," *British Journal of Industrial Relations*, 1984, *22*, 101–112.

Kremer, Michael, "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics*, August 1993, *108* (3), 551–575.

___ **and Eric Maskin**, "Segregation by Skill and the Rise in Inequality," February 1994.

Krueger, Alan B. and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, March 1988, *56* (2), 259–293.

Kuznets, Simon, "Economic Growth and Income Inequality," *American Economic Review*, March 1955, *45* (1), 1–28.

MacLeod, W. Bentley, "Contract, Complexity and the Employment Relationship," February 1996. Boston College.

___ **and James M. Malcomson**, "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica*, March 1989, *57* (2), 447–480.

___ **and** ___ , "Wage Premiums and Profit Maximization in Efficiency Wage Models," *European Economic Review*, August 1993, *37* (6), 1223–1249.

Millward, Neil, Mark Stevens, David Smart, and W. R. Hawes, *Workplace Industrial Relations in Transition: The ED/ESRC/PSI/ACAS Surveys*, Aldershot: Dartmouth, 1992.

Moene, Karl Ove, "A Reformulation of the Harris-Todaro Mechanism with Endogenous Wages," *Economics Letters*, 1988, *27* (4), 387–390.

Pearce, David G., "Repeated Games: Cooperation and Rationality," in Jean-Jacques Laffont, ed., *Advances in Economic Theory: Sixth World Congress*, Vol. I, Cambridge, UK: Cambridge University Press, 1992, chapter 4, pp. 132–174.

Posner, Richard A., *Economic Analysis of Law*, 3rd ed., Boston, MA: Little Brown, 1986.

Raff, Daniel M. G., "Wage Determination Theory and the Five-Dollar Day at Ford," *Journal of Economic History*, June 1988, *48* (2), 387–399.

____ **and Lawrence H. Summers**, "Did Henry Ford Pay Efficiency Wages?," *Journal of Labor Economics*, October 1987, *5* (4, part 2), S57–S86. Supplement.

Shapiro, Carl and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74* (3), 433–444.

Stewart, James B., "Taking the Dare," *The New Yorker*, July 23 1993, pp. 433–444.

Troske, Kenneth, "Evidence on the Employer Size Wage Premium from Work-Established Matched Data," Technical Report, U.S. Census Bureau, Washington, DC 1994.

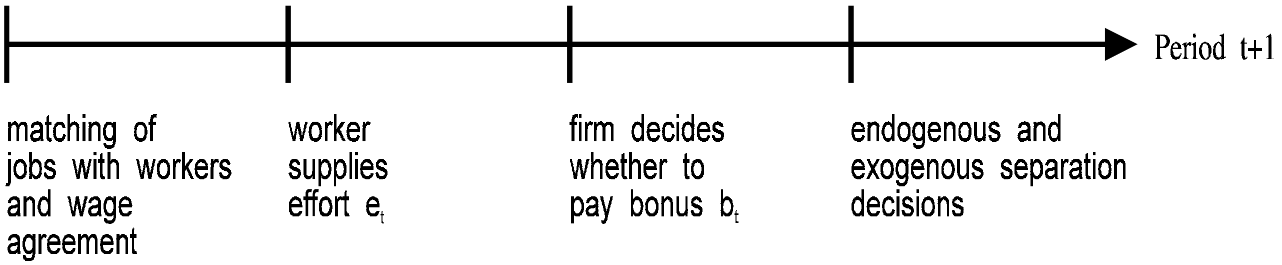
Period t+1
matching of jobs with workers and wage agreement
worker supplies effort e_t
firm decides whether to pay bonus b_t
endogenous and exogenous separation decisions

Figure 1: Timing of Events in Period t

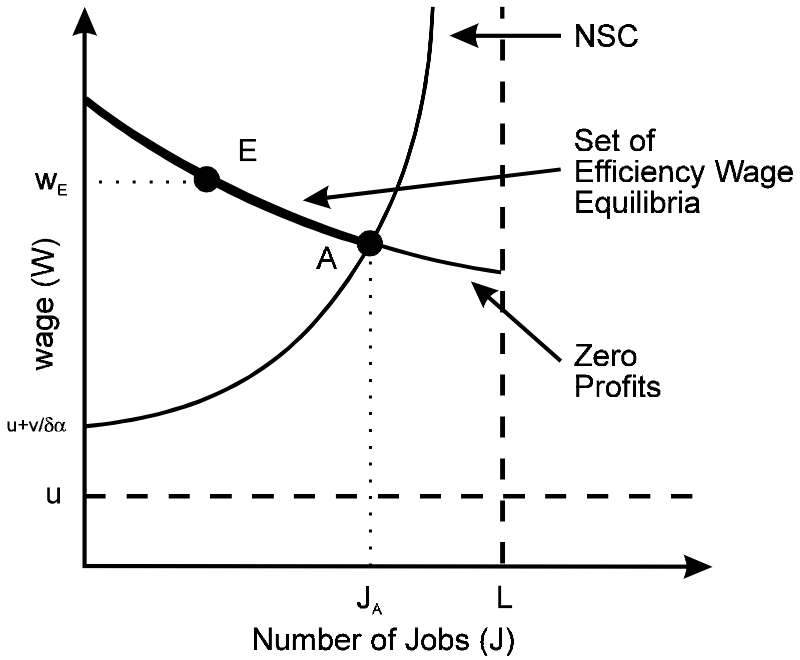

Figure 2: Market Equilibrium with Efficiency Wages

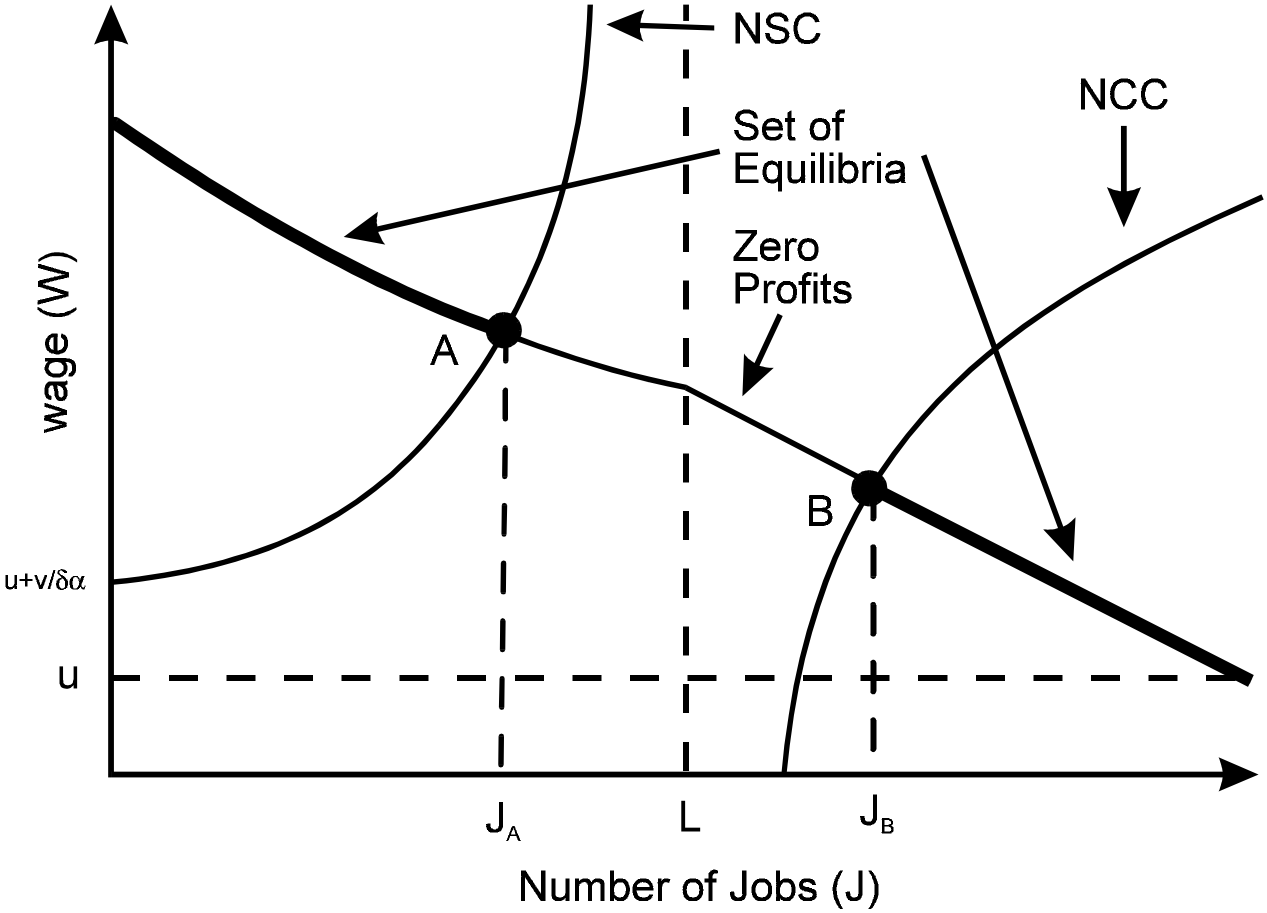

Figure 3: Market Equilibria with Performance Pay and Efficiency Wages

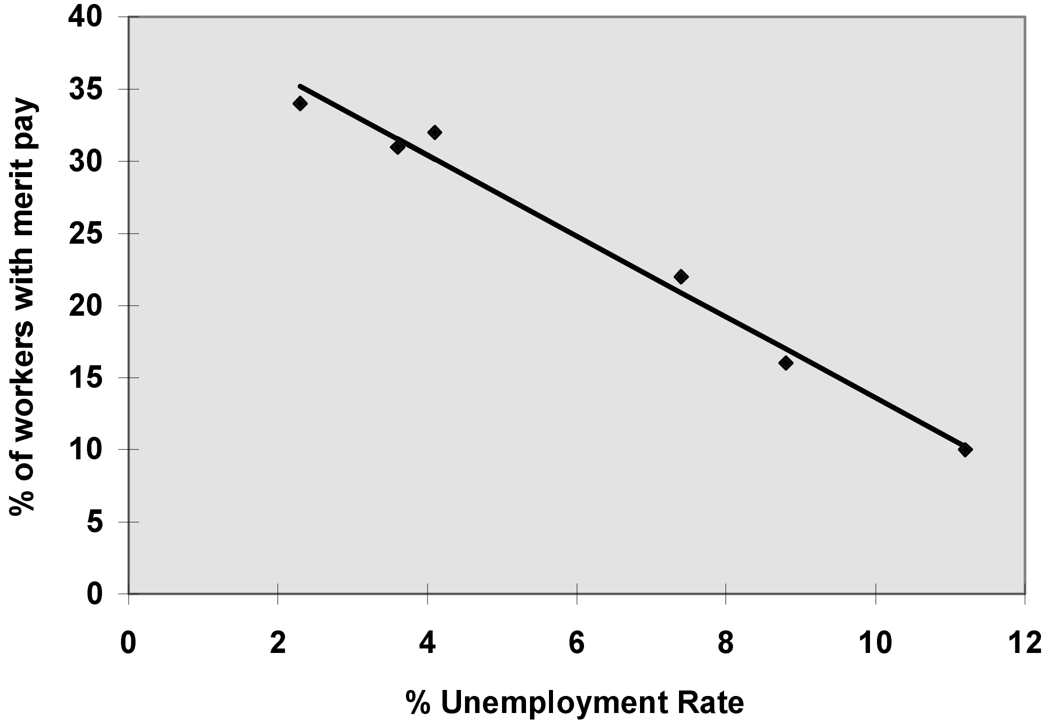

Figure 4: Percentage of Workers Receving Merit Pay versus Unemployment Rate by Occupational Group

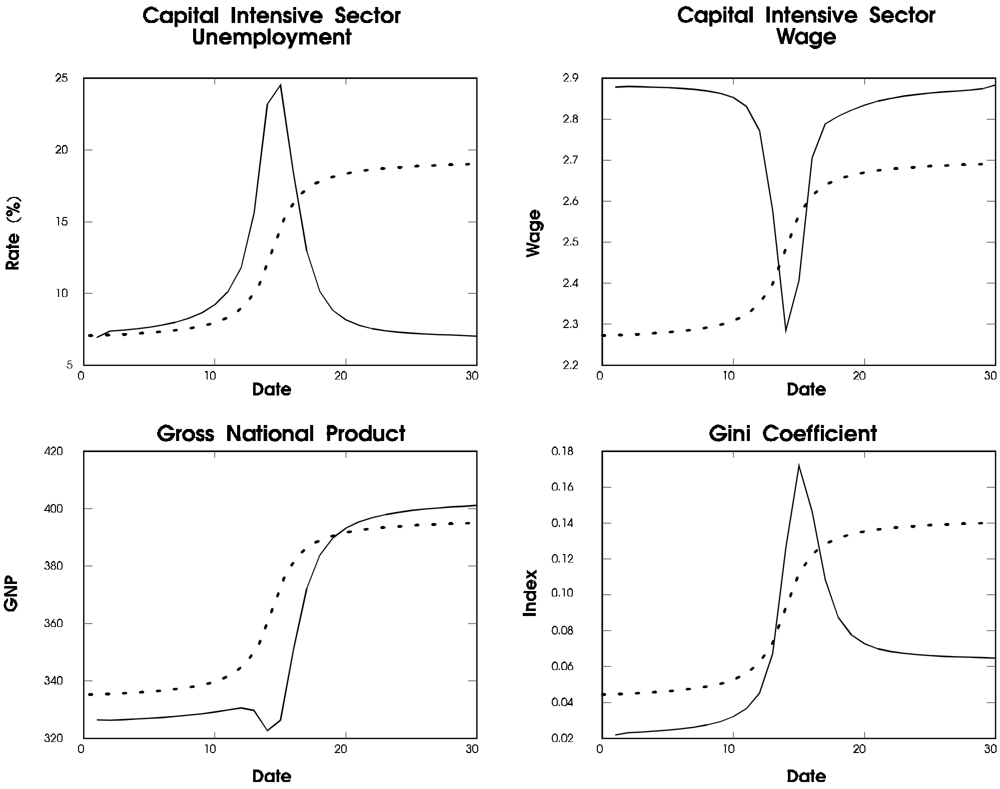

Figure 5: Economic Consequences of Growth