Fair Wages, Worker Motivation, and Implicit Bargaining Power in Segmented Labor Markets

by

WILLIAM D. FERGUSON*

This paper addresses implicit bargaining power within employment relationships using a versatile model of labor market segmentation that combines labor discipline, performance pay, insider power, and fair wage principles. In the primary sector, fair wage comparisons, firm-specific human capital, and less perfect monitoring engender bilateral bargaining power, yielding high compensation, sometimes including a bonus. Secondary-sector employers exert unilateral bargaining power, via credible dismissal threats with no replacement costs, and offer no bonuses. Differential determinates of implicit bargaining power can potentially explain various phenomena, including nominal wage rigidity, union wage differentials, job-specific wage differentials, and gender or race-based wage differentials. (JEL: J 40, J 30, J 50)

1 Introduction

This paper presents a model of labor market segmentation based on a combination of labor discipline, performance pay, insider power, and fair wage principles. The ensuing wage–motivation relationship operates in a context of implicit bargaining power, which functions differently in primary and secondary sectors of a dual labor market. Implicit bargaining need not be formal; it arises whenever one party can credibly threaten to impose sanctions on the other. Because employment relationships create dynamics of reciprocity and reside in an environment of imperfect contracting, employers and workers frequently possess an ability to threaten sanctions.

The primary sector of the dual labor market is characterized by “good” jobs with relatively high wages, favorable working conditions, relatively long-term employment tenure, along with the use of internal labor markets and relatively long

* In preparing this paper I benefited from research assistance from Wanlin Liu and Kate Anderson, comments from Paul McGrath, Raymond Robertson and Daniel Aaronson on early drafts, as well as more recent and very helpful comments from two anonymous referees.

Journal of Institutional and Theoretical Economics
JITE 161 (2005), 1–29 © 2005 Mohr Siebeck – ISSN 0932-4569
William D. Ferguson

promotion ladders (see DOERINGER AND PIORE [1971]). The secondary sector lacks these features, at least in a relative sense. In the primary sector, both workers and firms possess implicit bargaining power, reflecting, among other things, the acquisition of firm-specific human capital by workers – which renders them costly to replace – as well as fair wage comparisons among workers. In the secondary sector, by contrast, absence of these and other features allows employers to exert unilateral implicit bargaining power via credible dismissal threats, yielding lower wages. The secondary sector, moreover, does not offer performance pay, whereas the primary sector can when motivation conditions permit. Systematic differences in implicit bargaining power associated with underlying conditions in this model can, at least partially, explain a variety of phenomena including nominal wage rigidity, job-specific, firm or industry wage premia, and race and gender-based differentials.

This paper draws from a broad literature on incomplete contracting related to the problem of motivating effort in employment relationships. In this regard, standard efficiency wage models, or labor discipline models, indicate that firms pay wages above a market-clearing level in order to induce effort via potentially costly job loss (e.g., SHAPIRO AND STIGLITZ [1984]). A variety of authors proceed to apply these models to labor market segmentation. Some argue that the primary sector pays efficiency wages, whereas the secondary sector pays a market-clearing wage (e.g., BULOW AND SUMMERS [1986]).

In an innovative approach that allows for either efficiency wages or performance bonuses, MACLEOD AND MALCOMSON [1998] contend that incentive compatibility with regard to effort requires either an excess supply of or an excess demand for labor. Employers in a capital-intensive sector utilize jobs with high costs of vacancy. In order to rapidly fill openings, they desire an excess supply of labor, which is generated by their collective payment of efficiency wages, without performance pay. Firms induce effort in a manner that closely resembles the SHAPIRO AND STIGLITZ [1984] model. By contrast, employers in excess-demand labor markets use performance pay to induce effort.

Tracing further implications of incomplete contracting, BOWLES AND GINTIS [1990] develop a concept of power within market relations, which they call contested exchange. Here the party on the short side of a nonclearing market can exert power over the party on the long side, by threatening (perhaps implicitly) to terminate the relationship. Since termination is costly to the party on the long side, it imposes a sanction, which can alter behavior, indicating power. Referring to a labor discipline model, Bowles and Gintis maintain that employers exercise power

---

1 DICKENS AND LANG [1985a], [1985b] empirically distinguish the sectors on the basis of returns to human capital and experience, which are far stronger in the primary sector.

2 For MACLEOD AND MALCOMSON [1993], the secondary sector clears and the primary sector pays efficiency wages when markets are anonymous (previous dismissal history is unknown).

by implicitly threatening job loss in order to raise worker effort above the level which would obtain without such a sanction.

Applying contested exchange to the MacLeod–Malcomson model, an excess-supply labor market mirrors the above application. In an excess-demand (labor-intensive) sector, by contrast, workers occupy the short side of the market; they may credibly threaten to withhold effort in an environment where termination imposes costs on employers but not workers. To induce effort, then, employers must offer performance bonuses, reflecting worker power. In this formulation, neither party with power earns a rent; rather the possibility of denying rent to the opposite party constitutes the source of each party’s implicit bargaining power.

A contrasting treatment of power in labor markets appears in the insider-power models of Lindbeck and Snower [1986], [1988a], [1988b]. Here various forms of turnover costs, including training costs, search costs, and costs of noncooperation with new hires, confer market power to already employed workers (insiders), who then extract employment rents from their employer.

Finally, the fair wage literature (e.g., Akerlof [1982], Akerlof and Yellen [1990]) presents an alternative underlying mechanism for efficiency wages. Workers evaluate their current wages relative to other wages within applicable reference groups as well as their own previous wages. Groups of workers then establish norms regarding the appropriate or “fair” amount of effort to offer their employer; they apply motivated effort only when they feel their compensation is sufficiently fair. Knowing this, firms offer efficiency wages, a kind of “gift,” to motivate a reciprocal gift of effort above a minimum enforceable level.

Ernst Fehr and a variety of coauthors extend fair wage theory. Fehr, Gächter, and Kirchsteiger [1997] explicitly develop a conception of reciprocity: a willingness to reduce one’s own payoff in order to reward or punish behavior perceived as fair or unfair. Reciprocity, in turn, can serve as a means to enforce contracts and increase efficiency, by effectively discouraging “defect” options in relevant prisoners’ dilemma games. Proceeding further, Fehr and Schmidt [1999] develop an underlying concept of utility: some actors may possess “self-centered inequality aversion”: they suffer utility loss when their own payoff is excessively unequal to that of a relevant reference group, inducing a reciprocal punishment. Such reciprocity is likely to operate within the effort domain of employment relationships.

The present paper combines, extends, and alters elements of the above approaches. It develops a two-sector segmented labor market model for a developed (urban)
economy, founded largely on the MacLeod–Malcomson model, with important qualifications, caveats and additions. As in MacLeod and Malcomson, workers may exert two possible levels of effort, depending upon relevant incentives, and firms may motivate effort via dismissal threats or performance bonuses. The distinction between sectors, however, does not involve the presence of excess demand in a labor-intensive sector; rather, as the next section will detail, distinguishing primary-sector features include firm-specific human capital, less perfect monitoring, and relatively long-term employment tenure which facilitates fair wage comparisons. Indeed, unlike several treatments of segmentation with efficiency wages, here efficiency wages operate in both the primary and secondary sectors, implying an excess supply of labor in both.7

In the present model, the secondary sector functions like the capital-intensive sector in MacLeod and Malcomson [1998]: employers alone possess bargaining power, and firms pay efficiency wages with no bonuses. By contrast, the primary-sector model indicates bilateral implicit bargaining power, yielding higher compensation and, under certain conditions, a need or preference for offering limited performance bonuses. The decision to offer a bonus, only possible in the primary sector, does not depend on an excess demand for labor, but rather on interactions among factors such as the disutility and revenue gain derived from motivated effort, in relation to monitoring abilities of the firm.

Comparison of the two sectors, with attention to their underlying logic, yields a host of implications, including conditions that favor merit pay, sources of nominal wage rigidity, sources of union bargaining power, job, firm or industry wage premia, and race and gender wage differentials. Discussion will now turn to developing the primary sector model and proceed to distinguishing the sectors in the context of implicit bargaining power, followed by a discussion relating findings in relevant empirical literature, and finally a few applications.

2 Foundations and the Basic Primary-Sector Model

Following MacLeod and Malcomson [1998], we assume that (i) contracts which specify effort levels or criteria for performance bonuses are not enforceable in court, and (ii) capital-intensive firms face costs to opening and holding a job vacancy (c) that exceed necessary worker compensation. This last condition guarantees that capital-intensive firms pay efficiency wages in order to facilitate rapid filling of vacancies via an excess supply of labor. These two assumptions apply to both sectors of the dual labor market.8

7 Esfahani and Salehi-Isfahani [1989] develop a segmented labor market model with involuntary unemployment in both sectors. Note that an excess supply of labor exists in a one-sector efficiency wage model.

8 This model contains no labor-intensive sector in the sense of the rural sector at the end of MacLeod and Malcomson [1998]. With global capital mobility, a labor-intensive sector would appear in developing economies.
We proceed to elaborate the primary-sector model, which could stand alone as a one-sector fair wage/labor discipline, insider power model. Discussion of the secondary sector follows, in section 3. Several additional assumptions, not present in MacLeod and Malcomson [1998], distinguish the primary sector: (iii) because primary-sector workers develop firm-specific human capital, replacing them is costly. This assumption accords with Doeringer and Piore’s [1971] discussion of internal labor markets. Differences in sector production technology generate three additional assumptions: (iv) capital costs associated with creating a new vacancy, including costs of reorganization, are greater in the primary sector ($c_1 > c_2 > 0$; subscripts signify respective sectors); (v) the exogenous probability of termination of an employment relationship is lower in the primary sector ($1 - \rho_1 < 1 - \rho_2$; here $\rho$ signifies the probability of continued employment if both sides agree); (vi) because primary-sector skills are complex, firms monitor workers’ effort less perfectly than secondary-sector firms, who also monitor imperfectly ($d_1 < d_2 < 1$; $d$ indicates the probability of detecting a shirking worker). Here, task complexity and firm-specific human capital may interact: complex tasks may foster firm-specific human capital and interactions between complexity and specific human capital can further complicate monitoring.

Assumptions (iii), (iv), (v), and possibly (vi) generate relatively long employment tenure in the primary-sector. Such longevity is partly exogenous (from (v)) and partly endogenous: from (iii), primary-sector employers are reluctant to dismiss workers, but secondary-sector dismissal threats are always credible (p. 18). Assumption (vii) follows: because primary-sector workers enjoy relatively long-term employment, they compare their own remuneration with that of others in a relevant reference group. The utility gained from compensation, then, partially reflects such comparison. Note that assumption (iii) incorporates an element from insider-power models; assumption (iv) refines the MacLeod–Malcomson concept of capital-intensity to distinguish between the sectors; assumption (vi) adds imperfect information present in other labor discipline models, and assumption (vii) reflects fair wage considerations.

Turning to timing, as in MacLeod and Malcomson [1998], the decisions of a worker and firm in an already continuing employment relationship unfold in a current period ($t$) as follows: At the beginning of $t$, the firm offers the worker compensation ($W_t = w_t + b$, where $w > 0$, signifies the wage and $b \geq 0$ signifies a possible performance bonus). The worker decides whether to continue employment for the period and then whether to offer motivated effort ($e_M$) or low effort ($e_L$): $e_M > e_L \geq 0$. Effort should be construed broadly to include intensity and quality.

---


10 Malcomson [1999] notes that experimental studies imply that reciprocity increases with employment tenure. Campbell III and Kamlani [1997], find employers concerned about the perceived fairness of pay, relative to previous pay, pay among coworkers, and employer profits.
of work, propensity to innovate, etc. Normally, we expect $e_L > 0$, since idleness is boring. The firm then observes the worker’s effort, imperfectly, and decides whether to continue the relationship and whether to pay a performance bonus, if applicable. The period ends with the worker deciding whether to stay. Assuming no termination, subsequent periods repeat the same sequence.

With this foundation, we turn to the fair-wage utility function for a worker in a continuing employment relationship. Expected utility depends not only on traditional labor discipline factors – compensation, the disutility of additional effort, the probability of the relationship continuing, and the value of the outside option – it further reflects comparison of one’s own compensation with that awarded to others in the worker’s reference group. The latter can be expressed as a weighted average of own compensation and the difference between own and average reference group compensation: a comparison based on envy. The following utility function emerges:\[11\]

$$U_{t+1} = \alpha W_{t+1} + (1 - \alpha)(W_{t+1} - W_{ot}) - v + \delta[\rho_{t+1}U_{t+1} + (1 - \rho_{t+1})\bar{U}_{t+1}]$$

where $W, w, b, \rho$ remain as previously defined; $W_{ot}$ represents average total compensation for other workers in the reference group, which includes some more highly paid people (supervisors, managers) as well as previous values of the worker’s own wage; $v$ is the utility loss from offering motivated effort, the disutility of motivated effort; if $e = e_L$, $v = 0$; $\delta$ is the discount factor. We call $1 - \rho$ the probability of exogenous termination; $\rho_{t+1} < 1$. Following MacLeod and Malcolmson [1998], we assume that $\rho$ is independent of employment tenure. $\bar{U}_{t+1}$, the worker’s outside option, is the utility associated with being unemployed at the beginning of $t + 1$. Finally, $\alpha(0 < \alpha < 1)$ shows the utility weight placed on own compensation; the lower $\alpha$, the greater the envy of others’ wages. We expect low $\alpha$’s to occur in jobs with close contact and long tenure. The fair wage comparisons reflected in (1) are consistent with Bewley’s [1999] finding that employers are concerned about worker morale, which in turn responds to perceived fairness of pay, particularly with regard to internal wage comparisons.\[12\]

The expected profit for a primary-sector firm which intends to continue an employment relationship with a specific worker can be represented as follows:

$$\Pi_{t+1} = Q(e) - W_{t+1} + \delta \rho_{t+1} \Pi_{t+1} = q_e - w_{t+1} - b + \delta \rho_{t+1} \Pi_{t+1},$$

where $Q$ is the revenue product of labor, a function of $e$, and $q$ is technical productivity: output per unit effort (a function of employment, capital, and technology). If exogenous events terminate the relationship (with probability $1 - \rho_{t+1}$), $\Pi_{t+1} = 0$.

\[11\] We abstract from possible sectoral differences in $e, v, \delta, \mu$ (the unemployment benefit, see equation (3)). Note that $\alpha, b, W_{ot}$ do not apply to the secondary sector.

\[12\] Howitt [2002], surveying much relevant experimental and survey literature, reports a similar concern for fairness.
In equilibrium (see section on the worker motivation condition) workers have an incentive to apply motivated effort \((e_M)\).

2.1 Participation Constraints

In order for an employment relationship persist in \(t\), the worker’s employment utility and the firm’s profits must be at least as great as their respective outside options. The outside utility can be defined as

\[
\bar{U}_1 = \varepsilon_1(E_1)U_1 + (1 - \varepsilon_1(E_1))(\mu + \delta\bar{U}_{1t+1}),
\]

where \(\varepsilon_1\), the probability of attaining primary-sector employment, is a function of \(E_1\) actual primary-sector employment in a period such that \(\varepsilon_1 = (1 - \rho_1)E_1/(L_1 - \rho_1 E_1)\). Here \(L_1\) is the available primary-sector labor force; \(L_1 = L - L_2\), where \(L = L_1 + L_2\) is the total labor force, and \(E = E_1 + E_2\) is total employment; \(\mu\) is the unemployment benefit.\(^{13}\) The numerator indicates the rate of primary-sector job creation; the denominator shows the available labor force minus continuing employment. We assume workers can only search in one sector at a time and that competition among workers ensures that \(\bar{U}_1 = \bar{U}_2 = \bar{U}\) (see section 3). Because jobs in each sector are identical, the distinction between \(U_1\) and \(\bar{U}_1\) depends only on \(\varepsilon_1(E_1)\) and \(\mu\).

For the firm’s outside profit, its profit when it hires a new worker, we have

\[
\bar{\Pi}_1 = q_1e_M - W_1 + T + \delta\rho_{1t+1}(\Pi_{1t+1}),
\]

where \(T\) is the cost of training a replacement worker who lacks firm-specific human capital at the beginning of \(t\). In (3) and (4) we assume, for simplicity, that a new worker will receive the same compensation as the previously employed worker and that the employment relationship becomes a normal continuing relationship at the end of \(t\), with no subsequent training costs.\(^{14}\)

Given the above, the worker and firm participation constraints (WPC and FPC) depend upon the relationship between expected employment utility or profit and the outside options. More precisely:

\[
U_{1t} \geq \bar{U}_1 \quad \text{and} \quad \Pi_{1t} \geq \bar{\Pi}_1.
\]

To simplify further, following MacLeod and Malcolmson [1998], we assume a stationary equilibrium so that net job creation equals net job loss, both occurring at the constant probability of exogenous termination \((1 - \rho)\). Consequently, quits and dismissals indicate labor market churning; neither affects net job creation or

\(^{13}\) This specification arises from MacLeod and Malcolmson [1998, p. 396], explicitly adapted to two sectors.

\(^{14}\) MacLeod and Malcolmson [1998, p. 395], citing Kaufman [1984] and Bewley [1995], note that fair wage conventions often preclude paying new employees less than current employees. Using a one-period starting wage in the primary sector equal to the secondary-sector wage, does not significantly affect the results of this model.
loss. With stationarity, compensation and the revenue product grow at the same rate 
($\Pi_I = \Pi_{I+1}$; $\Pi_{I'} = \Pi_{I'+1}$, and $U_I = U_{I+1}$, and $U' = U_{I'+1}$).

As in other efficiency wage models, equilibrium employment falls short of the 
labor force ($E_I < L_I$; $E < L$). Consequently, worker utility depends on the wage 
package and the level of employment ($E$), as do employer profits. The utility and 
profit conditions under stationary equilibrium may now be rewritten as follows:

\begin{equation}
U_1(W_1, E_I) = W_1 - (1 - \alpha)W_o - v + \delta(\rho_1U_1 + (1 - \rho_1)\bar{U})
\end{equation}

(1a) \quad \frac{1 - \delta \rho_1}{1 - \delta} W_o - v + \delta(1 - \rho_1)\bar{U}

\begin{equation}
\Pi_I(W_1, E_I) = \frac{q_1\epsilon_M - W_1}{1 - \delta \rho_1} = \frac{q_1\epsilon_M - W_1}{1 - \delta}. \quad (2a)
\end{equation}

Similarly, market conditions influence the outside options for both firms and 
workers. We find:

\begin{equation}
\bar{U}_1(W_1, E_I) = \epsilon_1(E_I)U_1 + [1 - \epsilon_1(E_I)]\mu + \delta \bar{U}_1
\end{equation}

(3a) \quad \frac{\epsilon_1(E_I)U_1 + [1 - \epsilon_1(E_I)]\mu}{1 - \delta + \delta \epsilon_1(E_I)}

and

\begin{equation}
\Pi_I(W_1, E_I) = q_1\epsilon_M - W_1 - T + \delta \rho \Pi_I(W_1, E_I) = \Pi_I - T. \quad (4a)
\end{equation}

Equations (1a)–(4a) can be used to define the worker participation and employer 
participation constraints: $U_1(W_1, E_I) \geq \bar{U}(W_1, E_I); \Pi_I(W_1, E_I) \geq \Pi_I(W_1, E_I).$ After 
some manipulation, we find the worker participation constraint:

\begin{equation}
W_1 \geq (1 - \alpha)W_o + v + \mu. \quad (5a)
\end{equation}

The firm participation constraint reduces to

\begin{equation}
T \geq 0. \quad (6a)
\end{equation}

In primary-sector equilibrium, both equations hold as strict inequalities, indicating 
employment rents for both workers and firms.

Discussion now turns to three types of incentive compatibility conditions: (i) the 
worker motivation condition, which is the general condition for motivating 
effort; (ii) the dismissal credibility condition, the condition for motivating effort 
via dismissal threats alone, and (iii) the bonus incentive conditions, the conditions 
related to motivating effort with a bonus.

### 2.2 The Worker Motivation Condition

Assume that equations (5a) and (6a) hold. Consider now whether, having already 
received an agreement specifying $W$, the worker has an incentive to exert motivated 
effort in period $t$. We first compare the worker’s utility from exerting low effort 
(shirking) to her utility in equation (1). Define the shirking utility ($U_{1t}$) as

\begin{equation}
U_{1t} = \epsilon_1W_t - (1 - \alpha)W_o + (1 - d_t)(b_t + \delta \rho_{t+1}U_{1t+1} + \delta(1 - \rho_{t+1})\bar{U}_{t+1})
\end{equation}

(7) \quad + d_t \delta \bar{U}_{t+1}.
Equation (7) states that the utility associated with shirking depends on four factors: (i) the wage (paid anyway), (ii) the fair wage comparison (which occurs anyway), (iii) the utility gained if not detected—a shirking worker who is not detected in \( t \) will continue to shirk in \( t+1 \); and (iv) the expected utility if detected and dismissed: the discounted outside utility.\(^{15}\)

We begin with the general condition for motivating effort, under which both dismissal and performance bonuses are considered possible—for the moment abstracting from the issue of the credibility of dismissal threats or promises to pay bonuses. In order for the worker to exert motivated effort, the ensuing utility must be at least as great as the shirking utility (\( U_t \geq U_{st} \)). Using (1) and (7), we find:

\[
\delta \rho_{t+1} (U_{1t+1} - U_{1st+1}) + \delta \rho_{t+1} d_t (U_{1t+1} - \bar{U}_{t+1}) + d_t b_t \geq \nu. \tag{8}
\]

The sum of the expected future utility gain from motivated effort plus the expected difference between the shirking utility and the outside utility, in the event of dismissal, plus the expected increment in \( b \) which is induced by motivated effort must be at least as great as the disutility of motivated effort. Note further that the lower \( d_1 \), or the higher \( \nu \), the greater the required differences indicated by the first or second terms. We call (8) the worker motivation condition (the WMC).

With stationary equilibrium and some rearranging, we can redefine equations (7) and (8). The stationary utility from shirking becomes

\[
U_{st}(W_1, E_1) = w_1 - (1 - \alpha) W_o + (1 - d_1)b + \delta (1 - \rho_1 + d_\rho_1) \bar{U}
\]

\[
1 - \delta \rho_1 + d_\delta \rho_1.
\]

The worker motivation condition (WMC), after significant manipulation, can be expressed in terms of total compensation:

\[
W_1 \geq (1 - \alpha) W_o + \nu + \mu + \frac{1 - \delta \rho_1 + \varepsilon_1 \delta \rho_1}{d_\delta \rho_1 (1 - \varepsilon_1)} (v - d_1 b).
\tag{8a}
\]

Equation (8a) yields several underlying conditions for the motivation of effort via wages:\(^{16}\) The minimum compensation required to motivate effort (\( W \)) increases in (i) the unemployment benefit (\( \mu \)), (ii) the disutility of motivated effort (\( \nu \)), and (iii) the probability of attaining employment (\( \varepsilon \)), as long as \( d_1 b < \nu \); assume this for the moment. On the other hand, (iv) \( W \) decreases in the probability of detection (\( d_1 \)). Conditions (i)–(iii), applied to wages, are standard to the efficiency wage literature. The present model, however, allows for performance bonuses as part of total compensation.\(^{17}\) Condition (iv), again applied to wages, appears in some efficiency wage arguments. Continuing with implications, (v) \( W \) decreases in the expected value of

\(^{15}\) Assuming the worker was willing to risk shirking in \( t \), non-detection implies no change in circumstances for \( t+1 \) (the worker retained her job and even received her bonus). If the worker were unwilling to continue shirking in \( t+1 \), she would not have shirked in \( t \).

\(^{16}\) Derivation of (8a) is available upon request from the author. Equation (8a) is a complex version of the SHAPIRO AND STIGLITZ [1984] no shirking condition (NSC).

\(^{17}\) MACLEOD AND MALCOMSON [1993], [1998] explicitly consider performance bonuses as an alternative to efficiency wages: firms that pay efficiency wages cannot offer bonuses.
future employment, indicated by the product $\rho \delta$; (vi) $W$ increases in the comparison compensation ($W_c$). Condition (vi), applied to wages, reflects fair wage arguments, such as AKERLOF AND YELLEN [1990]. The present model combines these results from labor discipline and fair wage models and, by allowing bonuses, extends them to total compensation.

A result not present in other models: (vii) the motivation compensation decreases in the utility weight placed on the worker’s own compensation ($\alpha$). From (8a), we can see that $\partial W_1/\partial \alpha < 0$. Intuition: $\alpha$ reflects the utility weight placed on the worker’s own compensation independent of comparison; the smaller $\alpha$, the more important the comparison and the greater the utility loss from a given $W_0$ (envy). Accordingly, the worker’s utility increases in $\alpha$; the firm may motivate effort with a correspondingly lower $W$, ceteris paribus.

To summarize so far, conditions (i), (ii), (iii) and (iv) imply an employment rent, whereas conditions (vi) and (vii) affect the size of the worker’s reservation wage. The final implications of equation (8a) concern the payment of bonuses, generating the following proposition:

**Proposition 1** If both dismissal threats and bonus promises are valid, the payment of performance bonuses motivates effort more effectively than wages, yet the necessity of paying an employment rent limits the size of the bonus such that $db < v$.

From (8a), we find $\partial W_1/\partial b < 0$. Intuition: workers who are detected shirking fail to receive a bonus in the current period, yet they still receive the current period’s wage, even if fired at the end of the period. With respect to motivating effort, then, firms should prefer paying bonuses to paying wages, ceteris paribus, when feasible. Nevertheless, consistent with observations that firms rarely rely solely on bonuses and frequently offer no bonuses, the model limits the size of bonus payments. Recall, assumption (ii) indicates that the capital costs of vacancy ($c$) exceed the labor costs of an occupied position. Production efficiency, then, requires offering workers an employment rent via efficiency wages. The right hand side of (8a) indicates that the $W$ necessary for motivation must at least be equal to the $W$ required by the WPC (the right hand side of (5a)) plus the difference $v - db$ multiplied by its parameter. The necessity of offering an employment rent, therefore, limits the size of any bonus such that $db < v$.

Unlike most efficiency wage models, the presence of $T > 0$ allows efficiency-wage firms the potential to offer bonuses (subject to $db < v$). Examination of (5a), moreover, indicates that firms must pay a positive wage. In section 2.4 we consider bonus effectiveness and credibility conditions, which can impose additional restrictions on bonuses. Risk aversion among workers, along with institutional constraints and conventions may further restrict the value of feasible bonuses. In some important cases the maximum feasible bonus is zero.

---

18 This logic parallels that found in MACLEOD AND MALCOMSON [1998], but with no $T$, their efficiency wage sector pays no bonus.

19 FEHR AND GÄCHTER [1998] argue that prevalence of fixed wage contracts may arise because complex tasks render awarding individual merit difficult and explicit per-
Broadly speaking, equation (8a) indicates two mechanisms for inducing motivated effort, which might be characterized as carrot and stick: (i) the firm promises to award a bonus to workers not detected shirking, and (ii) it threatens, implicitly at least, to dismiss workers who are detected shirking. Given $T > 0$, the present model allows the possibility for simultaneous operation of both mechanisms in the primary sector. In order to consider their relative applicability, we proceed to analyze conditions that underlie the credibility or effectiveness of relevant threats or promises. We begin with the dismissal mechanism.

2.3 The Dismissal Credibility Condition

Because primary-sector firms face training costs, an implicit threat to dismiss a worker caught shirking may lack credibility; dismissal may cost the firm more than tolerating shirking. Such credibility depends on the relative profitability of dismissing a detected shirking worker compared to retaining such a worker. In this regard, define $\Pi_{sd}$ as the firm’s profit when it observes a worker shirking, withholds that worker’s bonus for the current period, and dismisses the worker at the end of the period. Assume that once disciplinary dismissal has actually been used on an individual, future occupants of her job will exert effort whenever equation (8a) holds.\(^{20}\) Accordingly, we have:

$$\Pi_{1sd} = q_1e_L - w_t + \delta \rho_{1t+1}(\Pi_{1t+1} - T_{t+1}).$$

(9)

If, on the other hand, a worker who shirks is detected and denied a bonus but not dismissed, we assume the worker will continue to shirk in the next period, provided denial of a bonus is not sufficient to deter shirking.\(^{21}\) We define the associated profit function ($\Pi_{st}$) as follows:

$$\Pi_{st} = q_1e_L - w_t + \delta \rho_{1t+1}\Pi_{st+1},$$

(10)

where $\Pi_{st+1} = q_1e_L - w_{t+1} - (1 - d_1)b + \delta \rho_{1t+2}\Pi_{st+2}$. Assuming stationary equilibrium, we have

$$\Pi_{sd} = q_1e_L - w_t + \delta \rho_{1t} \left[ qe_L - w_t + b \right] 1 - \frac{\delta \rho_{1t}}{1 - T_{t+1}},$$

(9a) and

$$\Pi_{st} = q_1e_L - w_t + \delta \rho_{1t} \left[ qe_L - w_t + (1 - d_1)b \right] 1 - \frac{\delta \rho_{1t}}{1 - T_{t+1}}.$$

(10a)

Credibility of a dismissal threat requires that $\Pi_{sd}$ be at least as great as $\Pi_{st}$. After some manipulation, we find:

$$q_1(e_L - e_t) \geq (1 - \delta \rho_{1t})T + d_1b.$$  

(11)

\(^{20}\) Performance incentives may reduce the trust necessary for cooperation. Biewley [1999] reports similar concerns.

\(^{21}\) Here we assume internal reputation effects, but no external reputation effects.
Equation (11) describes the dismissal credibility condition (the DCC), establishing the next proposition.

**Proposition 2 (the DCC)** In order for a firm’s dismissal threat to attain credibility, the value of the productivity gain from motivated effort must be at least as great as the sum of the expected cost of training a new worker and the value of any current bonus times the probability of detection.

Note here that performance bonuses and dismissal threats act as quasi substitutes: for a given degree of imperfect monitoring, the higher \(b\) (within limits of Proposition 1) the more difficult attaining equation (11) becomes.\(^{22}\) Finally, if both \(q\) and \(T\) depend on job, firm, or industry, the feasibility of motivating effort via threatened dismissal may differ across these categories.

### 2.4 Bonus Credibility and Effectiveness Conditions

Now consider the credibility of a firm’s promise to offer a performance bonus. In an excess-supply (efficiency wage) labor market, with \(T = 0\), such a promise is not credible; the employer would lose nothing if it failed to honor its promise and the worker quit. Here, however, firm-specific human capital renders \(T > 0\), generating employment rents for employers. Even so, bonus offers are not always credible.

As previously indicated, the firm pays an applicable bonus at the end of \(t\), after motivated effort has been expended (assuming (8a) holds). The firm, then, could renege on the bonus without losing the current period’s effort, but if it did, the worker would shirk in \(t + 1\). The firm, therefore, will honor its promise to pay the bonus if future motivated effort justifies the current bonus payment. In equation form:

\[
q_1e_M - w_1 - b + \delta\rho_1\Pi_{t+1} \geq q_1e_M - w_1 + \delta\rho_1\Pi_{t-1} \tag{12}
\]

where \(\Pi_{\text{nb}}\) indicates profits earned when a bonus is not paid for motivated effort, and \(\Pi_{\text{nb}t+1} = q_1e_L - w_1 + \delta\rho_1\Pi_{\text{nb}t+2}\). With stationarity, we have \(\Pi_{\text{nb}} = (qe_L - w_1)/(1 - \delta\rho_1)\). Using this and equation (3a), with some manipulation, we find the following intuitively appealing result:

\[
\delta\rho_1q_1(e_M - e_L) \geq b. \tag{13}
\]

Equation (13) establishes the bonus credibility condition (the BCC): the credibility of an offer to pay a bonus requires that the discounted revenue gain from future motivated effort be at least as great as the current bonus. Comparing (13) with (11), however, we see that as long as \(T > 0\), and both the DCC and Proposition 1 hold, the BCC is not a constraint.\(^{23}\)

\(^{22}\) If a shirking worker is fired and replaced, the firm is committed to paying a bonus indefinitely into the future, whereas if it retains the shirking worker, it pays no future bonus when shirking is detected.

\(^{23}\) Substituting the maximum possible \(b\) from (13) into (11), we find a less stringent constraint on \(q_1(e_M - e_L)\). Whenever the DCC holds, then, Proposition 1 sets the upper limit on \(b\) (assuming no outside constraints).
Now consider the case when the DCC fails: the firm cannot credibly threaten dismissal. Even if equation (8a) holds, payment of wages alone, or wages with a less than sufficient bonus, will no longer motivate high effort. The firm might still motivate effort by offering a sufficiently large performance bonus: workers found shirking will not be dismissed; they just fail to receive a bonus. Assuming bonus credibility, this strategy works only if $b$ is sufficiently large to induce motivated effort in the absence of a credible dismissal threat. Call this the bonus effectiveness condition, which constitutes the WMC when the DCC fails.

To proceed, assume that a shirking worker will not quit, if denied a bonus. Equation (7) must be rewritten so that the worker’s shirking utility ($U_{1s't}$) becomes:

$$U_{1s't} = w_1 - (1 - \alpha)W_o + (1 - d_1)b + \delta_1 U_{1s't+1} + \delta(1 - \rho_1)\bar{U}_{lt}.$$  

(14)

Assuming stationarity, the condition $U_t \geq U_{1s't}$ generates the bonus effectiveness condition:

$$d_1b \geq v \quad \text{or} \quad b \geq v/d_1.$$  

(15)

Equation (15) dictates the minimum effective bonus, the minimum bonus required to motivate effort when the DCC fails. Considering Proposition 1, however, we find:

**PROPOSITION 3** Whenever the vacancy costs associated with a position exceed its labor costs, firms cannot motivate effort by the payment of a bonus in the absence of credible dismissal threats.

From Proposition 1, $b < v/d_1$. Equation (15), therefore, cannot be attained: the WMC cannot be attained without credible dismissal threats.

Here the model duplicates the MacLeod and Malcomson [1998] result that dismissal threats must be present to motivate effort in a capital-intensive sector, yet it allows efficiency-wage firms limited bonus payments when and only when the DCC holds. Proposition 1 indicates that firms prefer performance bonuses to achieve (8a), yet the size of such bonuses is limited. Thus we expect both motivation mechanisms to operate for primary-sector firms where cultural and institutional constraints do not preclude performance bonuses.

Proceeding to tradeoffs, equation (11) indicates that the DCC depends upon the relationship between training costs and the value of motivated effort, and moreover, that payment of a bonus makes the DCC more difficult to attain. We can represent relevant incentive alternatives by graphing equation (11) and Proposition 1 in productivity/training cost space ($qTe$ space), as shown in Figure 1. We begin with the most easily attained DCC: the DCC0 curve shows the DCC for $b = 0$. Areas above and to the left of DCC0 indicate sufficient gain from motivated effort for

---

24 All else equal, equation (2) shows that the firm is always better off not paying a bonus, thus the implicit threat to withhold a bonus for insufficient effort is always credible.

25 Naturally, the WPC must hold. Intuition: if the worker knew she would shirk and then quit after not receiving the bonus, she never would have taken the job to begin with. This can be shown. In an excess-demand labor market, workers who quit lose nothing, but here quitting workers lose their employment rent.
William D. Ferguson

Figure 1
The DCC and Incentive Compatibility

\[ q(e_M - e_L) \geq (1 - \delta_p)T \]

Key: 
- Z - incentive compatible for any \( b < \nu/d_1 \)
- Y - not incentive compatible for \( b = b_A \)
- X - not incentive compatible for \( b = \nu/d_1 \)
- W - not incentive compatible

The DCC to hold: \( q(e_M - e_L) \geq (1 - \delta_p)T \); areas below and to the right are not incentive compatible (NIC). We know however that when possible, a firm prefers to offer a bonus to complement wages. If the firm offers a bonus level of \( b_A \), the DCC curve shifts up to DCC\(_A\), indicating a larger NIC area. Even so, if the firm gains enough from motivated effort, if \( q(e_M - e_L) \geq (1 - \delta_p)T + d_1 b_A \), DCC\(_A\) will hold, and the firm will offer a bonus at least as large as \( b_A \), if Proposition 1 and exogenous constraints allow. At the maximum possible bonus, we find a still larger NIC; the intercept of DCC\(_b\) on Figure 1 falls just below the point \( \nu/d_1 \). Overall, then, firms may offer performance bonuses only with sufficiently large gains to motivated effort; the larger the bonus, within the limits of Proposition 1, the larger the required gain.

In conclusion, the primary-sector model generates the following hypotheses regarding effort motivation and the payment of performance bonuses: (i) Given training costs, the more productive a job (firm), the more likely the attainment of motivated effort and the greater the likelihood that bonuses will complement wages. (ii) Given training costs, the longer expected employment duration of a given job,
the more likely the attainment of motivated effort and the greater the likelihood that bonuses will complement wages. (iii) The higher the level of firm-specific human capital generated by a job, the greater the productivity required to achieve motivated effort and the lower the likelihood that bonuses will complement wages. (iv) Jobs (firms) with easily observable effort are more likely to offer bonuses. (v) Jobs (firms) with a high disutility of effort are more likely to pay bonuses.

Hypothesis (i) follows because, given $T$, highly productive firms gain more from motivating effort than they might lose through dismissal, and, as Proposition 1 indicates, the DCC is a prerequisite to the payment of bonuses. Hypothesis (ii) employs the same logic, applied to the expected future value of a given employment relationship. Hypothesis (iii) follows because high training costs render dismissal less credible. Hypotheses (iv) and (v) follow from Proposition 1. Overall, productive jobs or firms with relatively low training costs are more likely to attain effort incentive compatibility. Consequently, they are more likely to offer bonuses, especially if effort is both easily observable and relatively distasteful.

A discussion of implicit bargaining power in the context of segmented labor markets generates additional implications.

3 Labor Market Segmentation and Implicit Bargaining Power

The secondary sector lacks the amenities of the primary sector; it pays relatively low wages for jobs with relatively short tenure. As indicated, assumptions (i) and (ii) apply to both sectors; consequently, both sectors employ efficiency wages and face an excess supply of labor. Referring back to assumptions (iii)–(vii), we formally distinguish the secondary sector from the primary sector as follows: there is no firm-specific human capital, hence $T_2 = 0$; sector production technology yields lower capital costs of vacancies ($0 < c_2 < c_1$), as well as a higher exogenous probability of termination ($1 - \rho_2 > 1 - \rho_1$), and a lesser degree of task complexity, and consequently less imperfect monitoring ($1 > d_2 > d_1$). Given the above, employment tenure is relatively short-term, allowing no fair wage comparisons ($\alpha = 1$).

We then distinguish secondary-sector values for the terms $W$, $U$, $\Pi$, $\varepsilon$, $E$, $d$, $c$, $\rho$, and $q$ in equations (1)–(8) via use of subscript 2. Furthermore, $b_2 = 0$, implying $W_2 = w_2$, and $\alpha(=1)$ drops from (1) and other relevant equations. With $T_2 = 0$, dismissal threats are always credible – the firm has nothing to lose – and bonus promises are never credible. Consequently, equations (9)–(15) do not apply to the secondary sector. The Appendix shows the full secondary-sector model. The most important equations, with some treatment of rationale and implications, follow.

As in Acemoglu [2001], we assume that unemployed workers may search in either sector, but only one sector at a time. Competition among the unemployed, then, equalizes the unemployment utilities for both sectors.26 At stationary equilibrium,

26 Esfahani and Salehi-Isfahani [1989] make the same assumption. Both papers find that equivalent outside utilities between the two sectors indicate longer job queues in the primary sector.
we find
\[ \hat{U}_2(w_2, E_2) = \varepsilon(E_2)U_2 + \frac{[1 - \varepsilon(E_2)][\mu + \delta\hat{U}_2]}{1 - \delta + \delta\varepsilon(E_2)} = \hat{U}_1(W_1, E_1). \] (3b)

With \( T_2 = 0 \), equations (2a) and (4a) simplify to
\[ \Pi_2(w_2, E_2) = \Pi_2(w_2, E_2) = \frac{q\varepsilon\mu - w_2}{1 - \delta \rho_2}. \] (4b)

Turning to motivating workers, with \( T_2 = 0 \) and \( E_2 < L_2 \), \( \alpha = 1 \), an always viable DCC, and no bonuses, the worker motivation condition (WMC) for the secondary sector becomes:
\[ w_2 \geq \mu + \nu + \frac{1 - \delta \rho_2 + \varepsilon(E_2)\delta \rho_2}{d_2\delta \rho_2(1 - \varepsilon(E_2))}. \] (8b)

MACLEOD AND MALCOMSON [1998] indicate that adherence to their WMC (NSC) requires a fair wage convention. In this regard, current effort in equation (8) depends on future rather than current compensation, since \( W \) is set at the beginning of the period. The firm could then reduce the current wage slightly without affecting current effort; the worker would still receive an employment rent and would not quit. Recognizing this possibility, however, the worker will not apply motivated effort if the expected wage falls below a specified level. The firm understands this, and the efficiency wage becomes a fair wage convention.

Here it is necessary to distinguish between fair wage conventions and individual fair wage comparisons. Conventions exist among groups of workers and focus on the relationship between compensation and expectations concerning effort of others (expected strategies of others). They are implicit in any version of equation (8), including (8b), but stronger in the primary sector. Individual fair wage comparisons, on the other hand, reside in the utility function of individual workers (equation (1)) and focuses on comparison with reference group wages. Only the primary sector possesses sufficient employment tenure to generate such comparisons.

3.1 Market Equilibrium
We begin with market equilibrium in the secondary sector. The WMC, equation (8b) indicates that \( w_2 \) is a positive, increasing function of secondary-sector employment \( (E_2) \). Turning to the demand side, if we now assume that competition drives overall profits to zero, the profit generated from an employment relationship is equal

---

27 Equation (8b) is equivalent to the NSC in MACLEOD AND MALCOMSON [1998, p. 397], adjusted for \( d_2 < 1 \).
28 Enforcing a fair wage convention implies reciprocity in the sense used by FEHR, GÄCHTER, AND KIRCHSTEIGER [1997]. MACLEOD AND MALCOMSON [1993] contend that fairness conventions influence beliefs about the strategies of others, enabling selection among multiple Nash equilibria in a repeated prisoners’ dilemma game concerning effort.
(2005)  

Fair Wages

...to the sunk cost of creating a new job vacancy \( (c_2) \). Explicitly including diminishing returns (to rising \( E \)) in \( Q \), we find \( \Pi_2 = (Q(e_2, E_2) - w_2)/(1 - \delta p_2) = c_2 \), or

\[
(16) \quad w_2 = Q(e_2, E_2) - (1 - \delta p_2)c_2.
\]

We call equation (16) the zero-profit condition. Market equilibrium for the secondary sector occurs at the intersection of lines representing (16) and (8b) drawn in \( W - E \) space, as shown in Figure 2. Equation (16), labeled ZPC, slopes downward; equation (8b), labeled WMC₂, slopes upward, asymptotic to the vertical line \( L_i \) where \( E_2 = L_2 \) and the denominator vanishes.

![Figure 2: Sector Equilibria](image)

Key: \( W_{F1} \) = primary-sector wage with strong fair-wage convention; \( W_1 \) = primary-sector wage without strong fair-wage convention; \( W_2 \) = secondary-sector wage.

Different possible fair wage conventions, however, allow for multiple equilibria, beginning at the intersection of ZPC with WMC₂ and extending left along ZPC.²⁹ If evolutionary efficiency arguments apply to competitive markets (as in HAYEK [1982]), the equilibrium may be forced, over time at least, to a minimal fair wage convention, at the intersection of these two curves (as in SHAPIRO AND STIGLITZ [1984]). For simplicity, we will assume this to be the case for the secondary sector.

²⁹ Wage-effort conventions establish the possibility of multiple cooperative solutions to a prisoners’ dilemma game concerning wages and effort, as in LEIBENSTEIN [1987].
Primary-sector equilibrium follows the same principles, adjusted for sectoral differences. Here, WMC1 (8a) indicates a similar positive, increasing relationship between \( W_1 \) and \( E_1 \). The profit condition for a primary-sector firm with a continuing employment relationship appears in equation (16), with subscripts 1 replacing 2, but now (16) reflects the presence of an employment rent, arising from avoidance of \( T_1 \), at equilibrium. We call primary-sector version of equation (16) the firm rent curve (FRC1). As shown in Figure 2, FRC1 lies above the ZPC, which reflects conditions for entering primary-sector firms or firms with new vacancies.\(^{30}\) Strong fair wage conventions may push the fair wage above and to the left of the intersection of WMC1 and FRC1 (see point \( W_F \)). These last two points have implications for our next topic.

3.2 Implicit Bargaining Power

To proceed, we employ a simple definition of bargaining power: the ability of one party to credibly threaten sanctions on another party.\(^{31}\) Recall, bargaining here is implicit, signifying that bargaining power can exist without formal mechanisms: unions are not required. Unions formalize bargaining and clearly affect implicit bargaining power, but their presence is not a prerequisite for its existence. Turning to bargaining power within the sectors, we find that secondary-sector employers exercise unilateral power, utilizing credible threats of potential job loss as a sanction to motivate effort. In the primary sector, by contrast, training costs render power bilateral. We develop four related propositions, beginning with the secondary sector.

**Proposition 4** Secondary-sector employers possess unilateral implicit bargaining power over workers, but the degree of such power at a given wage decreases in the level of employment and the unemployment utility.

In the secondary sector, employers can replace workers without cost and the DCC always obtains. The ensuing potential threat of job loss induces workers to provide effort above the voluntary level (\( e_M \) instead of \( e_L \)), indicating employer power. Because firms may replace workers without cost, workers possess no similar sanction. The degree of employer power, however, depends on labor market conditions. If an exogenous shock, such as a monetary expansion, increases employment, the firm must increase its wage in order to maintain a sufficient job-loss sanction, as reflected in the upward slope of the WMC2 curve. Hence as \( E_2 \) increases, the firm’s bargaining power at a given wage diminishes. Similarly, an increase in \( \mu \) would decrease

\(^{30}\) For drawing simplicity, both sectors share the same ZPC in Figure 2. Zero profits apply to primary-sector firms with new employment relationships. For such firms, the profit equation is \( W_1 = Q(e_1, E_1) - T - (1 - \delta_D)\epsilon_1 = 0 \), implying \( \Pi_1 = \epsilon_1 \). Firms with continuing employment relationships, need not pay \( T \), hence their employment rent.

\(^{31}\) BACHARACH AND LAWLER [1981] use a similar definition: one party’s bargaining power emerges from the degree of the other’s dependence on it.
the cost of job loss at a given wage, shifting \( WMC_2 \) up, indicating a higher \( w_2 \) at any given \( E_2 \).

By contrast, the distinguishing features of the primary sector limit employer bargaining power and introduce implicit bargaining power for workers. For employers, we have

**Proposition 5** The implicit bargaining power of primary-sector employers decreases in the level of employment and the unemployment utility, and is further mediated by imperfect monitoring and fair wage comparisons.

As in the secondary sector, exogenous labor market changes (affecting \( E \) or \( \mu \)) condition employers’ ability to impose sanctions, generating a shift or movement along the increasingly steep \( WMC_2 \) curve. Less perfect monitoring, however, weakens the firm’s bargaining position by rendering disciplinary dismissal less certain. Because \( d_1 < d_2 \), the employment rent indicated by \( WMC_1 \) exceeds that from \( WMC_2 \), reflecting a reduction in the firm’s bargaining power for any given \( W \). Fair wage comparisons further mediate employer bargaining power. An increase in either the utility weight of the comparison wage \( (1 - \alpha) \) or the comparison wage itself \( (W_0) \) will increase the worker’s reservation wage (see equation (5a)), reducing the potential job-loss sanction for any given \( W \). These last two effects push \( WMC_1 \), in Figure 2, above \( WMC_2 \), indicating a higher \( W \), *ceteris paribus*.

Turning to primary-sector workers, we have

**Proposition 6** The presence of firm-specific human capital and relatively strong fair wage conventions in the primary sector generate worker implicit bargaining power.

Firm-specific human capital renders primary-sector workers costly to replace \( (T > 0) \). As reflected in equation (6), training costs generate employment rents for firms with continuing employment relationships. Possible denial of such rents confers implicit bargaining power on primary-sector workers. Accordingly, in Figure 2, \( FRC_1 \) lies above \( ZPC \), pushing \( W_1 \) further above \( W_2 \). Further, as suggested in sections 3 and 3.1, any fair wage convention above a minimal level pushes the fair wage above \( W_1 \) \( (W_F \text{ in Figure 2}) \). This augmentation of \( W \) reflects enhanced worker implicit bargaining power because the implicit potential to withhold motivated effort, arising from possible violation of a fair wage convention, offers workers a potential sanction.

---

32 The firm derives no utility from power. If it could achieve \( e_M \) without a sanction, it would. If effort were a continuous function of the cost of job loss, as in Bowles and Gintis [1990], exogenous factors that increase the cost of job loss would also increase effort at a given wage.

33 Power here resembles extraction of appropriable quasi-rents (Klein, Crawford, and Alchian [1978]). A rise in \( T \) increases the distance between \( ZPC \) and \( FRC_1 \), signifying implicit worker bargaining power. By contrast, an increase in marginal productivity would shift both curves upwards.
Turning now to bonuses, the previous discussion indicates that bonus offers are not credible unless $T > 0$: bonus credibility exists only if workers possess some implicit bargaining power. The possibility of withholding a bonus, however, increases the employer’s implicit bargaining power by creating another possible sanction, shifting $WMC_1$ down and to the right. The final proposition follows:

**Proposition 7** Where feasible, primary-sector firms will use bonuses to enhance their own implicit bargaining power, facilitating lower total compensation.

The model may now be extended.

4 Extensions

We may augment the model with additional turnover costs, specifically costs of search for employees and noncooperation by insiders when outsiders are hired. Both factors potentially enhance the implicit bargaining power of primary-sector workers. Concerning search, if primary-sector firms need a one-period search to locate suitable employees for any opening, turnover costs rise. The distance between the ZPC and $FRC_1$ increases by one period’s value of motivated effort ($q e_M$); $W_1$ rises accordingly (not shown in Figure 2).\(^{34}\)

Concerning noncooperation, assume that following the dismissal of a worker, perhaps in violation of a convention, some group (size $g$) of remaining workers (insiders) may decide not to cooperate with a new hire (an outsider) by offering only $e_L$. The $FRC$ would shift further to the right, reflecting the potential loss of value from motivated effort ($g q(e_M - e_L)$). Furthermore, restoring motivated effort in subsequent periods may require paying the $g$ workers a higher wage; they may have adjusted their fair wage convention upwards. Determination of the number $g$, however, involves a complex collective-action problem since the firm’s ability to retaliate via dismissal or denying bonuses may limit willingness to withhold cooperation.\(^{35}\) Still, whenever $g > 0$, the $FRC$ lies further to the right and $W_T$ may lie higher than it otherwise would, both effects reflecting enhanced implicit bargaining power of workers.

We may further augment the model by adding a craft sector, where occupational supply restrictions generate excess demand for craft labor.\(^{36}\) Here the DCC never holds, but competition among firms may require a relatively high $W$ to meet the $WPC$ (5a), and motivation can be attained only by paying a large enough bonus. Equation (15) now specifies the minimum effective bonus ($b_*$), indicating the relevant $WMC$.\(^{36}\)

\(^{34}\) Applying search costs to (6), we find the applicable rent: $\Pi_1 - \bar{\Pi}_1 = T + q e_M$.

\(^{35}\) Workers and firms may engage in negative reciprocity, as in **FEHR, GÄCHTER, AND KIRCHSTEIGER [1997]**, indicating a willingness to suffer costs in order to punish “unfair” behavior. Precise determination of $g$ lies beyond the scope of this paper.

\(^{36}\) Craft rules limit access to relevant markets **OSTERMAN [1988]**, **MACLEOD AND MALCOMSON [1998]** cite sports stars and financial traders as examples of workers in short supply; both fit the present definition of craft workers.
Combining (15) with bonus credibility (13), craft-sector firms can motivate effort whenever \( q(e_M - e_L) \geq v/dP \): in the absence of a credible dismissal threat, a bonus strategy becomes both credible and effective only if the value of the productivity gain from motivated effort is large enough relative to the disutility of effort, accounting for imperfect monitoring and discounting. Institutional constraints, which can specify a maximum bonus, may prevent certain jobs, firms, or industries from entering the craft sector by precluding the attainment of \( b_{\text{c}} \). Further, the requirement of paying a sufficient bonus to motivate craft-sector effort reflects relatively weak employer bargaining power, since dismissal threats always lack credibility in the craft sector.

5 Evidence in the Literature

The present model is consistent with findings in a large number of empirical studies which report experimental, survey, and econometric evidence, as well as stylized wage patterns. A sizeable body of experimental and survey data point to the importance of motivation and fairness in employment relationships. MALCOMSON [1999], reviewing this literature, offers three important findings: (i) support for the notion that workers have scope with respect to effort, indicating that motivation matters, (ii) evidence that both workers and firms participate in reciprocal behavior, (iii) evidence for the existence of employment rents. He adds that surveys report employer resistance to paying new hires market-clearing wages for fear of negative impacts on morale. The current model indicates similar resistance to wage cuts.

Turning to econometric evidence of efficiency wage and fair wage effects, KRUEGER [1991] finds that company-owned restaurants, which have less incentive than franchises to closely monitor employees, pay higher wages to managers and slightly higher wages to crew workers. He also finds higher wage-tenure profiles in franchises, consistent with efficiency wages, and notes that the relatively short tenure in fast food tends to support efficiency wage theory. Similarly, REBITZER [1995] finds a negative tradeoff between wages and supervision in the petrochemical industry. ARAI [1994], using Swedish data, finds evidence that “autonomy” can help explain inter-industry wage differentials, in a manner consistent with efficiency wage models. Summarizing relevant studies, MALCOMSON [1999] finds “considerable evidence” for the existence of employee rents, and some evidence that high wages reduce disciplinary dismissals. The present model indicates rents generated by turnover costs, monitoring costs, and the need to attain cooperation.

Concerning observed wage patterns, the model’s predictions are consistent with three key patterns noted by Malcomson and others. First, as the model predicts, there is evidence of firm or job-specific wage premia. For example, BROWN AND MEDOFF [1989] find that wages increase with firm size, noting that efficiency wages offer one

---

37 The reported differentials are 9% for managers and 2% for crew workers. While the latter figure is low, fast food workers occupy low rungs in secondary labor markets, where efficiency wage effects should be small, and minimum wage laws may constrain wage differentials.
of six possible explanations for this relation. Rebitzer and Robinson [1991] find that both plant and firm size increase primary-sector wages. Commenting on the need for a theory of employer rent sharing, they contend that plant-size effects may reflect primary-sector monitoring problems, as in Bulow and Summers [1986], but firm-size effects may instead reflect rent-sharing and reciprocity considerations. The present model’s incorporation of training costs, imperfect monitoring, fair wage comparisons, with implicit bargaining power, all of which may differ systematically by job, establishment, firm, or industry, offer a possible foundation for such a theory. Moreover, the use of efficiency wages in the secondary sector accords with Rebitzer and Robinson’s observation of secondary-sector plant-size effects. Finally, Malcomson [1999] states that models which allow for distinctions in monitoring costs, turnover costs, and rent sharing can explain firm or job-specific premia. The present model facilitates all three distinctions.

Secondly, Malcomson notes that wages, especially those for long-term employees, are less responsive to business cycles than spot models would predict. MacLeod, Malcomson, and Gomme [1994] find that a Shapiro–Stiglitz efficiency wage model can account for changes in real wages and unemployment over the business cycle. The present model’s incorporation of efficiency wages is consistent with these observations; moreover, its incorporation fair wage comparisons, with previous wages included in \( W_o \), enhances this dynamic by making it consistent with the notion of wage aspirations (see Blanchard and Katz [1997]).

Third, in the short run, nominal wages are somewhat rigid. Howitt [2002] contends that the nominal rigidity reflects employer fear that workers will reciprocate “unfair” nominal wages cuts with low motivation. If one were to interpret the comparison wage \( W_o \) in equation (1) to include the worker’s own previous nominal wage, the model would predict this kind of downward nominal wage rigidity, and would also strengthen a wage aspiration effect.

Overall, the model’s simultaneous incorporation of efficiency wages, fair wages, and training costs is consistent with all three stylized wage patterns. By merging and extending concepts from other models, the present model maintains substantial consistency with empirical evidence reported in the literature.

6 Applications

There are many potential applications of this model, ranging from the distribution of labor income to macroeconomic relationships between aggregate employment and...
wages. The previous discussion of stylized facts suggests that it may help explain nominal wage rigidity, aspiration wages, cyclical variations in unemployment, and job-specific pay premia. Due to space considerations, we proceed only with a few additional comments on job, firm, or industry premia, followed by a brief discussion of union bargaining power and then race and gender wage differentials.

The present model allows for and potentially reconciles contending explanations for job, firm, or industry wage premia found in relevant literature. Wage premia may reflect monitoring differences (as in BULOW AND SUMMERS [1986]), different capital costs (MACLEOD AND MALCOMSON [1998]), or different training or search costs. GIBBONS AND KATZ [1992], however, observe that the correlation of industry premia across occupations cannot be explained by monitoring differences. The present model’s inclusion of fair wage comparisons and conventions, operable within firms or establishments, offers a potential explanation. Furthermore, motivating effort in capital-intensive jobs, firms, or industries is likely to generate relatively large value differences \( q(e_M - e_L) \), requiring relatively high efficiency wages. Finally, some of the person-specific effects reported by ABOWD, KRAMARZ, AND MARGOLIS [1999] and ABOWD AND KRAMARZ [2000] (see note 38) may reflect identification of persons with specific jobs which systematically display differences in implicit bargaining power.39

6.1 Union Bargaining Power

According to LINDBECK AND SNOWER [2001], insider power models explain why unions can attain bargaining power: firms often find it less expensive to pay union wages than to face various turnover costs. The present model pursues this argument with respect to both training and search costs, and adds a fair wage component. As indicated in section 4, groups of workers \( g \) may decide to withhold motivated effort, either in response to hiring of outsiders or a failure to pay “fair” wages. Nonunion workers, however, may face difficult collective-action problems associated with attaining a large \( g \). Unions facilitate collective action, substantially enhancing insider power, realizing unutilized potential implicit bargaining power. In addition, obviously, unions add formal institutional sources of power, such as grievance procedures, the potential for strikes, and regular negotiations over wages and working conditions. Finally, the extended model’s implication of weak employer bargaining power in the craft sector may help explain the notable historical success of unions in craft occupations.

6.2 Race and Gender-Based Differentials

The concept of implicit bargaining power offers insight into wage discrimination by race or gender in the presence of profit-maximizing behavior. One theory of discrimination asserts that occupational crowding or occupational segregation limits

---

39 This interpretation appears to be consistent with MALCOMSON’s [1999] discussion of job wage premia.
the access of women and/or minority workers to specific (primary-sector) jobs and thereby “crowds” them into other, usually secondary-sector, jobs (ANKER [1997], BERGMANN [1986], KING [1992], MACPHERSON AND HIRSCH [1995], TOMASKOVIC-DEVRY [1993]). If so, white male workers in the primary-sector can face relatively high reemployment probabilities and lower costs of job loss than their nonwhite or female counterparts, ceteris paribus. Consequently, the WMC\(_1\) for white males may lie above and to the left of that for others, indicating less employer bargaining power vis-à-vis white males. Turning to search costs in the extended model, women and minority workers who attain access to primary-sector jobs would be easily replaceable, with relatively low search costs indicating correspondingly low levels of implicit bargaining power, hence a lower \(W\).

GREEN AND MCINTOSH [1998] present evidence that is consistent with this view of racial discrimination. Using data from the Workplace Industrial Relations Survey for Britain, they analyze the relationship between worker effort and a host of variables including the cost of job loss defined in a manner which is consistent with that used here. While they do not focus their analysis on labor market discrimination per se, they find high levels of “ethnic content” associated with relatively high effort, ceteris paribus (p. 378). They note that this result accords with efficiency wage theory if ethnic workers encounter a disadvantage in labor markets: facing a higher cost of job loss, ethnic workers will offer more effort to avoid losing their jobs.

The possibility of distinct fair wage comparisons and fair wage conventions presents another potentially important mechanism for the persistence of race or gender-based differentials. Suppose now, as BECKER [1971] asserts, that some white male workers dislike association with either women or nonwhite workers, requiring a compensating wage payment. Now consider the fair wage comparison in equation (1). Any such distaste for association would reduce the utility value placed on one’s own compensation (\(\alpha\) falls) and the disutility of \(W_o\) increases (envy may become resentment), driving up the white-male reservation wage. The WMC curve, accordingly, shifts up and to the left, reflecting reduced employer bargaining power over affected white-male workers, generating a higher \(W\), all else equal. Note that this specification allows for subtle worker prejudice, which may not occur at a conscious level.

Moving further, suppose that desire for distance not only affects one’s individual utility function, but influences fair wage conventions and effort expectations among groups of workers. If a firm, which previously had no or few women or minority workers in certain strategic (e.g., supervisory) positions, proceeds to add nonwhite or female workers to these positions, without boosting the wages of relevant white males, it risks noncooperation among certain prejudiced white males. Prevalence of prejudice or discriminatory habits may, moreover, solve collective-action problems associated with attaining a large \(g\); white male implicit bargaining power may become significant. To avoid noncooperation, the firm may still hire nonwhite male workers and increase the pay of currently employed white males, or it may segregate employment, perhaps by occupation, to preclude or at least reduce such comparison.
Because all three of these race or gender effects involve implicit bargaining power and its impact on the distribution of shared rents, market competition may not erode any resulting discrimination, as it does for Becker.

7 Conclusion

Implicit bargaining power resides in market transactions whenever it is costly for at least one party to leave a market relationship; such implicit power becomes bilateral whenever both parties face costs to exiting their relationship. This paper employs labor discipline, performance pay, insider power, and fair wage principles to generate a model of implicit bargaining power within a segmented labor market. In the secondary sector of such a market, termination of employment imposes costs on workers but not on firms; employers exert unilateral power over workers, yet still pay them efficiency wages. In the primary sector – where monitoring is less perfect; firm-specific human capital generates training costs; relatively long-term tenure facilitates fair wage comparisons, and fair wage conventions are stronger – implicit bargaining power is bilateral. Consequently, primary-sector firms pay higher efficiency wages and may offer bonuses to help motivate effort and temper worker bargaining power.

The model generates a complex set of implications regarding incentive compatibility of effort and its relationship to efficiency wages and performance bonuses. Secondary-sector firms can always motivate effort via credible dismissal threats, but cannot credibly offer bonuses. By contrast, primary-sector firms must have a large enough gain from motivated effort, relative to training costs, to achieve credible dismissal threats, the prerequisite for motivation. Such firms prefer to complement efficiency wages with performance bonuses; yet to remain credible larger bonuses require correspondingly larger gains from motivated effort, and, moreover, bonuses cannot be too large relative to the disutility of motivated effort (accounting for the probability of dismissal). Cultural and institutional factors may set additional limits. Productive primary-sector firms with low levels of firm-specific human capital, easily observed effort, and a high disutility of effort, then, are more likely than others to utilize bonuses. Further, in a craft sector of an extended model, bonuses are a prerequisite for motivation, reflecting relatively weak employer bargaining power. Still to justify bonuses, gains from motivated effort must be large enough relative to effort disutility, given dismissal probability and discounting. More productive firms have an advantage.

The model’s premises and workings accord with many empirical studies which report experimental, survey, and econometric evidence. It is also consistent with stylized observations concerning wage behavior over business cycles and nominal wage rigidity. Further, the concept of implicit bargaining power facilitates explanation of various wage differentials.

Concerning job-specific, firm, or industry wage premia, the model suggests that systematic differences in the determinates of implicit bargaining power can gener-
ate higher compensation in jobs where monitoring is difficult, firm-specific human capital is important, capital costs are high, fairness comparisons and fair wage conventions are strong, and worker cooperation is important. Resulting wage premia may then extend beyond those generated by human capital differences or compensating wage differentials, and market competition need not erode them. Concerning unions, the model finds that training and search costs, along with potential withholding of cooperation – if conventions are violated – and the ensuing potential for collective action underlie and reinforce union bargaining power. Finally, implicit bargaining power can differ systematically by race and gender, reflecting different employment probabilities, different search costs, as well as race or gender-oriented perceptions manifested through individual fair wage comparisons and fair wage conventions. Race and gender-based differentials, therefore, can persist without erosion by market competition.

Generally speaking, the model suggests that policy makers should pay close attention to determinates of implicit bargaining power and its underlying relationship with the dynamics of worker motivation. Those who strive, for example, to increase wages of the working poor, ought to address worker motivation and (lack of) implicit bargaining power in the secondary sector. They might ask: is it possible to achieve a Pareto superior outcome by altering constraints or incentives faced by firms and workers, perhaps facilitating higher compensation in exchange for higher motivation? Similarly, those who hope to reduce discrimination in labor markets need to address systematic differentials in implicit bargaining power which may exist among different race, gender, or other categories of workers and their relation to on-site conventions. Many other policy implications are possible; this model strives to offer a framework that may facilitate future analysis.

**Appendix: The Secondary Sector Model**

\[
U_2(w_2, E_2) = \frac{w_2 - v + \delta(1 - \rho_2)\bar{U}_2}{1 - \delta\rho_2},
\]

\[
\Pi_2(w_2, E_2) = \frac{q_2e_1 - w_2}{1 - \delta\rho_2},
\]

\[
\bar{U}_2(w_2, E_2) = e_2(E_2)U_2 + [1 - e_2(E_2)](\mu + \delta\bar{U}_2)
\]

\[
= \frac{e_2(E_2)U + [1 - e_2(E_2)]\mu}{1 - \delta + \delta e_2(E_2)}.
\]

\[
\bar{\Pi}_2(w_2, E_2) = \Pi_2(w_2, E_2) = \frac{q_2e_2M - w_2}{1 - \delta\rho_2},
\]

\[
w_2 \geq v + \mu,
\]

\[
T_2 = 0,
\]
\begin{align}
U_2(w_2, E_2) &= \frac{w + \delta(1 - \rho + d\rho)\bar{U}}{1 - \delta\rho + d\delta\rho}, \\
(A7b) \\
\frac{w_2 \geq \mu + v}{\frac{1 - \delta\rho + \epsilon(E_2)\delta\rho}{d\rho(1 - \epsilon(E_2))}v}. \\
(A8b)
\end{align}

References

– – – – – –, AND D. N. MARGOLIS [1999], “High Wage Workers and High Wage Firms,” Econometrica, 67, 251–333.
– – [1999], Why Wages Don’t Fall, Harvard University Press: Cambridge, MA.


William D. Ferguson
Department of Economics
Grinnell College
1210 Park St., D-3
Grinnell, IA 50112
USA
E-mail:
ferguso1@grinnell.edu