WORKER MOTIVATION, WAGES, AND BILATERAL MARKET POWER IN NONUNION LABOR MARKETS

William D. Ferguson
Grinnell College

INTRODUCTION

The terms of market exchange can, under certain conditions, even in competitive markets, reflect underlying power relationships. One important source of such power resides in the ability of either party to an exchange credibly to threaten to impose costs on the other party via termination of the relationship. Labor markets lend themselves to this kind of market power because employment relationships often constitute a positive sum game. Either or both parties may suffer from termination of the relationship.

Bilateral market power exists in employment relationships whenever both workers and employers benefit from maintaining their association: whenever the value of the employment relationship exceeds the value of the next best alternative for both parties simultaneously. Two conditions lead to the presence of bilateral power in employment relationships: i) job loss should be costly for workers; and ii) replacing workers should be costly for firms. Both conditions emerge whenever workers possess firm-specific human capital—or, more generally, whenever firms face costs to replacing current employees—and workers face reemployment costs.

This paper develops a widely applicable model of bilateral market power in employment relationships along with some of its implications. The model combines concepts from two strands of the efficiency wage literature—labor discipline and labor turnover models—with concepts from insider power models. It indicates that systematic differences in the cost of job loss to workers and the cost of replacing workers to firms may contribute to explaining a wide variety of observed patterns of wage differentiation—ranging from interindustry or race- and gender-based wage differentials to impacts of technical change and international competition on the skill distribution of labor income, as well as regional and macroeconomic wage patterns. The model's implications here are broadly consistent with results obtained in other portions of the literature. Moreover, by explicitly addressing the role of bilateral market power in wage determination, it adds explanatory dimension, allowing consistency with a broader range of phenomena than many other models.

William D. Ferguson: Economics Department, 1210 Park Street, D-3, Grinnell College, Grinnell, Iowa 50112-1670. E-mail: Ferguso1@grinnell.edu.

Eastern Economic Journal, Vol. 30, No. 4, Fall 2004

Literature Review

The efficiency wage and insider power literatures each offer potential conceptual foundations for a theory of market power in labor markets, typically presenting either condition i) or condition ii), though generally not both simultaneously. Matching models offer some additional perspective.

Within the efficiency wage literature, labor discipline (or effort) models, for example, Shapiro and Stiglitz [1984], Bowles [1985], and Bulow and Summers [1986], offer a convincing rationale for the presence of condition i) in employment relationships. Because the labor contract cannot fully specify the intensity and quality of worker effort applied within the space of an hour, employers must design an internal mechanism to ensure adequate effort. The effort dimension of the employment relationship, then, becomes a principal-agent problem.[1] The employer (the principal) strives to induce a profit-maximizing level of effort from the worker (the agent), whose level of effort, in turn, reflects individual utility maximization. The employer ensures appropriate effort by monitoring performance and offering a wage above the market-clearing level, indicating a positive cost of job loss which, in turn, gives the worker an incentive to exert herself on the job.[2]

Bowles and Gintis [1990] use a similar labor discipline model to illustrate the exercise of power in market exchange, reflecting what they call contested exchange.[3] In contested exchange, the party on the short side of a nonclearing market can exert power over the party on the long side by threatening (implicitly or explicitly) to terminate the relationship without fear of retaliation. Since termination is costly to the party on the long side, it imposes a sanction. Because threatened sanctions can alter behavior, they indicate power. In the case of a labor discipline model, employers use the implicit threat of costly job loss to induce workers to increase their effort above the level they would offer without such a sanction [Bowles and Gintis, 1990, 177-87].[4] The labor discipline model's solution to the effort problem, then, establishes condition i) for the existence of market power in employment relationships without condition ii). Employers possess unilateral market power because firms can implicitly threaten workers with costly job loss without risking worker sanctions.

A second kind of efficiency wage model, the labor turnover model [Salop, 1979], approaches the employment relationship from a nearly opposite angle, modeling condition ii) without much attention to i). For Salop, turnover from quits is a negative function of the wage paid; therefore, up to some point, firms can economize on turnover costs by raising wages. Even so, workers can face a positive cost of job loss—Salop's measure of labor market tightness implies costly job loss, and he even refers to firms' monopsony power in their internal labor markets—but the cost of job loss plays no role other than inhibiting worker quits.

The insider power models of Lindbeck and Snower [1988] take the implications of turnover costs several steps further. They argue that current employees with a certain degree of job tenure (insiders) are costly to replace, due to hiring and training costs and possibly a failure among colleagues to cooperate with new hires. Insiders, therefore, possess market power, which they use to attain employment rents. Like Salop, Lindbeck and Snower place little emphasis on the potential market power of the employer.

A related literature on matching models utilizes replacement costs as a factor in the simultaneous search of employers and workers for employment "matches." Hosios [1990], for example, examines the role of search costs, unemployment, and vacancies in a matching process to evaluate the efficiency of labor market outcomes. Resources devoted to employment search, by both employers and workers, affect the probability of attaining a productive employment match, defined as a match where resulting productivity is at least as great as the sum of reservation values and search costs. Turnover, then, may impose costs on both parties. While bargaining determines the distribution of any production surplus arising from successful matches, it has no effect on the matching process itself.[5] The Hosios model and other matching models imply a kind of bilateral market power in employment relationships, though they usually focus on other results, such as market efficiency or aggregate outcomes like the Beveridge curve.[6]

Cahill [2000] offers a related line of reasoning, combining a turnover cost model that exhibits matching principles with an efficiency wage model. For Cahill, efficiency-wage considerations determine a minimum acceptable wage for inducing effort (a no-shirking condition), whereas turnover costs determine a maximum profitable wage that allows firms to hold employment positions open. The intersection of the two resulting curves (graphed against employment), under the assumption of a steady-state matching process, determines equilibrium wage and employment levels. Cahill proceeds to argue that efficiency wage considerations (such as imperfect monitoring) push the wage upwards, whereas turnover costs reduce it.

The present approach resembles Cahill [2000] by merging efficiency wage and turnover cost arguments, but differs critically in its application of turnover costs. Whereas Cahill regards turnover costs as a constraint on wage setting via the profitability of holding vacancies open for potential new employees (outsiders), the current paper, following Lindbeck and Snower [1988], links turnover costs to insider power. Firms that retain insiders avoid turnover costs, earning them employment rents from continuing employment relationships. Because replacement costs offer insiders a potential sanction, they possess market power, allowing them to capture some of the firm's employment rent.

The present model, then, combines efficiency wage and turnover cost arguments in a fashion that focuses on the bilateral nature of market power in employment relationships. Efficiency wage considerations of labor discipline operate via an employer sanction of costly job loss, and turnover costs imply a sanction available to employed workers. Both conditions for bilateral market power are thereby achieved. Analysis now begins with the construction of a labor discipline model with turnover costs. Discussion then turns to applications of the model in six areas: overall relations between employers and individual workers, segmented labor markets, interindustry wage differentials, rising wage inequality, race and gender differentials, and a quick sketch concerning potential impacts of macroeconomic and regional implications.

A LABOR DISCIPLINE MODEL WITH DISMISSAL COSTS

As indicated above, this model combines the effort-inducement problem of a standard labor discipline (effort, or shirking) model with turnover costs to employers. Like

other labor discipline models, it employs standard competitive assumptions with three important qualifications, reflecting stylized facts of the labor process: 1) the amount of work actually performed cannot be fully specified *a priori* by contract; 2) the worker prefers to work at a level of effort that, in the absence of monitoring, falls below that which would maximize a firm's profits; and 3) it is costly for the firm to monitor worker performance.[7] These assumptions generate the agency problem associated with effort inducement common to effort models.[8] Following Bowles [1985] and Bowles and Gintis [1990], however, the present model adds two assumptions on effort that distinguish it from the Shapiro and Stiglitz [1984] model: 4) the effort function is continuous rather than dichotomous; 5) the worker's voluntary level of effort is positive, not zero.[9]

In order to solve its effort problem, the firm resorts to a strategy of contingent renewal: the worker may continue employment only if performance is deemed satisfactory. Contingent renewal depends upon costly monitoring of a worker's effort combined with an implicit threat of dismissal, under conditions where job loss is costly for the worker. So far, this logic reflects standard efficiency wage arguments. The present model, however, adds a final, complicating, assumption: 6) the firm faces costs of hiring and training any new worker. Such replacement costs complicate the relationship among wages, monitoring, and effort, forcing the firm to respond to its potential losses should it wish to dismiss a worker for disciplinary reasons. Given that voluntary effort is positive, high replacement costs may render toleration of low effort a less costly strategy than dismissal. Market power in employment relationships then becomes bilateral.

To proceed, assume a continuing employment relationship with the following time framework. At the beginning of the current period, the firm offers the worker a wage. The worker then decides her average level of effort for the period ($\hat{e}$). During the period, actual effort (e) varies stochastically around $\hat{e}$ (explained below), and the firm monitors actual effort periodically. At the end of the period, the firm considers its information on effort and decides whether to retain the worker. Future periods follow the same timing sequence.

Turning now to the model itself, we begin by developing the firm's dismissal function in several steps. Assume that the probability of dismissing a worker during a given time period depends on three factors: 1) the number of times the worker will be monitored during the period (M); 2) the adequacy of monitored effort, determined by comparing actual effort (e) at the moment observed with a work standard (α); and 3) the probability that the firm will actually fire a worker who has been monitored (F). Note that the firm chooses the monitoring input (m), along with α and F, subject to a variety of costs and constraints that will become apparent in the equations that follow. The worker chooses her average level of effort ($\hat{e}$); actual effort at any moment (e) is subject to random variation.

Each of these factors requires some explanation. The frequency of monitoring—the number of times a worker is observed during a period (M)—depends on the amount of monitoring input employed (m); a positive linear relationship is assumed.[10] Monitoring input, in turn, is a positive function of the firm's work standard (α), which specifies the minimum acceptable quality and intensity of effort. We assume here that

a higher α requires a higher m for enforcement, and that m is a negative function of the hourly cost of monitoring (μ). The following monitoring function ensues:

$$M = \lambda m(\alpha; \mu), \tag{1}$$

where the parameter λ reflects monitoring technology and a variety of working conditions that may either facilitate or hinder monitoring. We refer to λ as monitoring technology.

Now consider the second dismissal factor, the adequacy of effort. The firm chooses its work standard (α) within constraints set by available technology, industry, and occupation. We assume that once chosen, α remains fixed for the relevant time period. Recall that effort should be interpreted broadly to signify the intensity and quality of work executed during a specific hour. The worker chooses her average level of effort ($\hat{e}$), but actual effort at any point in time (e) is subject to random variation, arising from a variety of factors including the worker's health, fatigue, interruptions, etc. We assume that $e \sim N(\hat{e}, \sigma)$, where σ is exogenous and fixed. The difference between the firm's work standard and the observed level of effort ($\alpha - e$), then, determines the "adequacy" of the worker's effort: whenever $e < \alpha$, effort is "inadequate" and the worker may face dismissal.

Finally, concerning the third dismissal factor, we assume that the probability of firing a worker who has been monitored (F) depends upon the difference ($\alpha - e$) as well as expected hourly replacement costs arising from hiring and training a replacement worker (H). We assume that H is exogenous to the firm, influenced by industry, occupation, technical, and human capital factors.[11] Random fluctuation in e guarantees that periodically $F > 0$. More precisely, F depends on the product of the probability that $e < \alpha$, for a given ($\hat{e}$), and the ratio of H over the revenue product (P) of the difference ($\alpha - e$). We find

$$F(\alpha, \hat{e}; H, P) = \Pr(e < \alpha)\big|_{\hat{e}}\left[1 - H/P(\alpha - e)\right], \text{ for } P(\alpha - e) > H;\ 0 \text{ otherwise.} \tag{2}$$

Here $\Pr(e < \alpha)$ signifies the *a priori* probability. Naturally, $\partial F/\partial \hat{e} < 0$, $\partial F/\partial \alpha > 0$, and $\partial F/\partial H < 0$.[12]

Note the influence of replacement costs here: whenever $H = 0$, the second term in the brackets vanishes: all workers observed with $e < \alpha$ are fired. As H rises, however, dismissal becomes more costly, reducing F. Intuitively, if it costs the firm more to replace a recalcitrant worker than it does to keep the worker, the firm will refrain from dismissing the worker.[13]

Combining arguments from Equations (1) and (2), we find the dismissal function:

$$D = D(\alpha, \hat{e}; \mu, H, P) = M * F = \lambda m(\alpha; \mu) * \Pr(e < \alpha)\big|_{\hat{e}}\left[1 - H/P(\alpha - e)\right]. \tag{3}$$

D is the probability that a worker will be dismissed during a given period. D rises with α and falls with $\hat{e}$, μ, and H. Finally, if we were to simplify the model by assuming no replacement costs, the term H in Equations (2) and (3) would vanish. The model so construed would approximate a standard effort model.[14] Later we refer to the full model as model A, and the model without replacement costs as model B.

Given the employer's strategy of contingent renewal, with its attendant probability of dismissal from Equation (3), the worker chooses her average effort to maximize the value of the current employment relationship over the present and indefinite future (v). Adding the present employment utility (u) to the discounted value of possible future employment or unemployment, multiplied by the probabilities of retention and dismissal, respectively, we find:

$$v = u(\hat{e};w) + \left\{\left[1 - D(\alpha,\hat{e};\mu,H,P)\right]v + D(\alpha,\hat{e};\mu,H,P)z(b,U)\right\}\Big/(1+\rho)$$

Here ρ is the worker's discount rate and z is the worker's unemployment utility, a function of unemployment income (b) and the unemployment rate (U). With slight manipulation, we find:[15]

$$v = \left[u(1+\rho) + Dz\right]\Big/(D+\rho) = \left[u(1+\rho) - \rho z\right]\Big/(D+\rho) + z. \tag{4}$$

To proceed, we need to specify the current employment utility (u) and the unemployment utility (z). Utilizing assumptions 2) and 5), current employment utility is

$$u(\hat{e};w) = w - (\hat{e} - \bar{e})^2/2, \tag{5}$$

where $\bar{e}$ signifies the worker's voluntary level of effort, which is positive but less than the level of effort desired by the employer.

Concerning the unemployment utility (z), assume the following search process: after one period of search, the worker will find employment with probability $(1 - U)$. The employment utility, becomes

$$z = b + \delta\left(Uz + (1-U)v\right) = \left(b + \delta(1-U)v\right)\Big/(1-\delta U), \tag{6}$$

where δ is the discount factor $(1/(1 + \rho))$.

Incorporating Equations (3), (5), and (6) into (4), after some manipulation we find

$$v = \frac{u(1+\rho)(1-\delta U) + Db}{D(1-\delta) + \rho(1-\delta U)} = \frac{\left[w - \dfrac{(\hat{e} - e^2)}{2}\right](1+\rho)(1-\delta U) + \lambda mFb}{\lambda mF(1-\delta) + \rho(1-\delta U)}. \tag{7}$$

The first version of Equation (7) indicates that the value of current employment depends on the utility of current employment, the dismissal probability, the unemployment rate, unemployment income, and the discount rate. Specifying the dismissal and utility functions yields the second version of Equation (7). On the one hand, we find $\partial v/\partial m < 0$ and $\partial v/\partial \alpha < 0$, since an increase in either m or α increases the probability of dismissal from current employment (D). On the other hand, $\partial v/\partial \mu > 0$ and $\partial v/\partial H > 0$, since both reduce ($D$).[16] Equation (7) and its antecedents indicate further the dual role of effort: increasing $\hat{e}$ reduces the worker's employment utility (Equation 5), yet it also reduces the probability of dismissal (Equation 3). We now examine this topic in more detail.

Consider the worker's utility maximization problem over ($\hat{e}$) in light of its dual role. The first order condition with respect to average effort (setting $\partial v/\partial \hat{e} = 0$, in

Equation (4) or (7)) reveals that the worker chooses $\hat{e}$ in order to equate the marginal disutility of additional effort with the marginal value gain from a reduced probability of dismissal:

$$(8) \qquad \partial u/\partial \hat{e} - (\partial D/\partial \hat{e})\delta(v-z) = 0; \qquad or \qquad \partial u/\partial \hat{e} = (\partial D/\partial \hat{e}) * \delta(v-z).$$

Note that the term $\delta(v-z)$ is the expected cost to the worker of losing her job at the end of the current employment period.[17]

Equation (8) shows the worker's utility-maximizing response to the variables embodied in Equations (3)–(6) and ultimately (7). Moreover, by applying the implicit function theorem to Equation (8), we can specify the worker's choice of $\hat{e}$ in response to the variables embedded in these equations. We attain the following effort function in general form:[18]

$$(9) \qquad \hat{e} = \hat{e}(D,z,v) = \hat{e}(w,\alpha;\mu,H,P,U,b).$$

Using Equation (8), with its antecedents as the foundation of Equation (9), we attain several common efficiency wage results. A rise in the wage increases the cost of job loss, $\delta(v-z)$, raising the value of additional effort, presenting the worker an incentive to work harder, thereby increasing $\hat{e}$.[19] We assume further that $(\partial^2\hat{e}/\partial w^2) < 0$, reflecting diminishing returns of effort to increases in the wage. Turning to monitoring, Equations (1) and (3) indicate that an increase in μ reduces both m *and* M, reducing D, leading to a drop in $\hat{e}$, as long as the utility gain from reduced monitoring is not too large. Specifically, for a rise in μ to decrease $\hat{e}$, the absolute value of the cost elasticity of m times the cost of job loss must exceed the discounted net utility gain from the ensuing drop in D, for a given $\hat{e}$:[20]

$$\left|\frac{\partial m/\partial \mu}{m}\delta(v-z)\right| > \delta\left\{\frac{\partial(v-z)}{\partial D}(\partial D/\partial \mu)\right\}\Bigg|_{\hat{e}}.$$

The left-hand term shows the absolute value of the drop in effort incentive induced by a rise in μ; the right-hand term shows the gain in employment utility from the ensuing drop in D at a given $\hat{e}$.

In an analogous fashion, a rise in the work standard (α) makes dismissal for a given $\hat{e}$ more likely via its effects on m and F, again increasing $\hat{e}$, as long as the negative effect on employment utility is not too large. We find that $\hat{e}$ rises if:[21]

$$\left[\frac{\partial m/\partial \alpha}{m} + (\partial F/\partial \alpha)\Big|_{\hat{e}}\right]\delta(v-z) > \left|\delta\left[\frac{\partial(v-z)}{\partial D}(\partial D/\partial \alpha)\right]\Bigg|_{\hat{e}}\right|.$$

Here the left-hand term shows the increase in effort-incentive induced by a rise in D via α, and the right-hand term shows the absolute value of the loss in employment utility from the ensuing rise in D, at a given $\hat{e}$. The above results for μ and α suggest that prudent firms will limit the intensity of monitoring as they respond to a change in its costs as well as the severity of the work standard. For the remainder of the paper, we will assume both conditions hold.

Finally, in a comparable fashion, the unemployment benefit (b) and the unemployment rate (U) influence $\hat{e}$ by altering the relative attractiveness of employment versus the unemployment utility (z). Either an increase in b or a reduction in U will decrease the cost of job loss, reducing the incentive for exertion; $\hat{e}$ falls, *ceteris paribus*.[22] Similarly, an increase in H will decrease D, reducing $\hat{e}$, *ceteris paribus*.[23] Note that these last effects link the present model to macroeconomic or local labor market conditions. We will return to these links later.

To proceed further, we turn to the firm's profit maximization decision, given the worker's effort function. Profit, naturally, is revenue minus costs, shown as follows:

$$\pi = p^{*}\left[\chi^{*}\hat{e}(w,\alpha;\mu,H,P,U,b)\right]h-(w+\mu m+DH)h, \tag{10}$$

where π = total profits, net of capital costs; p = the market price of the firm's output; χ = output per unit effort (determined exogenously by capital stock and technology); and h = total labor hours employed by the firm. Assuming that the effort function, Equation (9), is independent of χ and h, the firm maximizes profits by setting w, m, α, and F to minimize the sum of total hourly costs per unit effort.[24]

In this regard, the wage, the monitoring input, the work standard, and the probability of firing serve as joint inputs into the production of worker effort. Thus, for given levels of χ and h, the firm minimizes its hourly labor costs per unit effort (Ψ):

$$\text{Min}\,\Psi=(w+\mu m+DH)/\hat{e}=\left(w+\mu m(\alpha;\mu)+\lambda m(\alpha;\mu)^{*}F(\alpha,\hat{e};H,P)H\right)/\hat{e}. \tag{11}$$

Note that H plays a dual role here: it is a cost to be minimized, and an increase in H reduces F for any given probability of finding $e<\alpha$.

From here, the firm's first order condition for wage setting follows:

$$\partial\hat{e}/\partial w=\hat{e}/\left[w+\mu m+\lambda mHF-\lambda mH\hat{e}(\partial F/\partial\hat{e})\right]. \tag{12}$$

The firm sets its wage so that the induced gain in $\hat{e}$ from a dollar wage increase equals the ratio of $\hat{e}$ to hourly labor costs, which, in turn, consist of w, μm, and expected hourly dismissal costs (the last two terms).[25] In a similar fashion, the firm chooses its monitoring input, responding to its costs and its relative effectiveness in generating effort. The first order condition for m is

$$\partial\hat{e}/\partial m=\left[\hat{e}(\mu+\lambda FH)\right]/\left[w+\mu m+\lambda mHF-\lambda mH\hat{e}(\partial F/\partial\hat{e})\right]=(\partial\hat{e}/\partial w)(\mu+\lambda FH). \tag{13}$$

The firm employs m up to the point where the induced effort gain equals the product of $(\partial\hat{e}/\partial w)$ and the sum of monitoring costs (μ) and λFH.

The firm chooses its work standard (α) with a similar comparison in mind. The first order condition with respect to the work standard is:

$$\begin{aligned}\partial\hat{e}\,/\,\partial\alpha&=\frac{\hat{e}\left(\mu(\partial m/\partial\alpha)+\lambda(\partial m/\partial\alpha)FH+\lambda mH(\partial F/\partial\alpha)\big|_{\hat{e}}\right)}{w+\mu m+\lambda mHF-\lambda mH\hat{e}(\partial F/\partial\hat{e})}\\&=(\partial e/\partial w)\left[\mu(\partial m/\partial\alpha)+\lambda FH(\partial m/\partial\alpha)+\lambda mH(\partial F/\partial\alpha)\big|_{\hat{e}}\right].\end{aligned} \tag{14}$$

Finally, for a given work standard, the firm's choice of the probability of firing (F) is simpler:

$$(\partial \hat{e}/\partial F) = \hat{e}\lambda mH/(w + \mu m + \lambda mFH). \tag{15}$$

To interpret Equations (13)–(15), it is instructive first to simplify the model by assuming no replacement costs ($H = 0$). In this case, model B, Equation (15) vanishes and Equations (12), (13), and (14) simplify to:

$$\partial \hat{e}/\partial w = \hat{e}/(w + \mu m)\,. \tag{12'}$$

For the monitoring input, we have

$$\partial \hat{e}/\partial m = \mu \hat{e}/(w + \mu m) = \mu(\partial \hat{e}/\partial w)\,. \tag{13'}$$

For the work standard, we have

$$\partial \hat{e}/\partial \alpha = \left(\mu(\partial m/\partial \alpha)\hat{e}\right)/(w + \mu m) = (\partial \hat{e}/\partial m)(\partial m/\partial \alpha) = \mu(\partial m/\partial \alpha)(\partial \hat{e}/\partial w). \tag{14'}$$

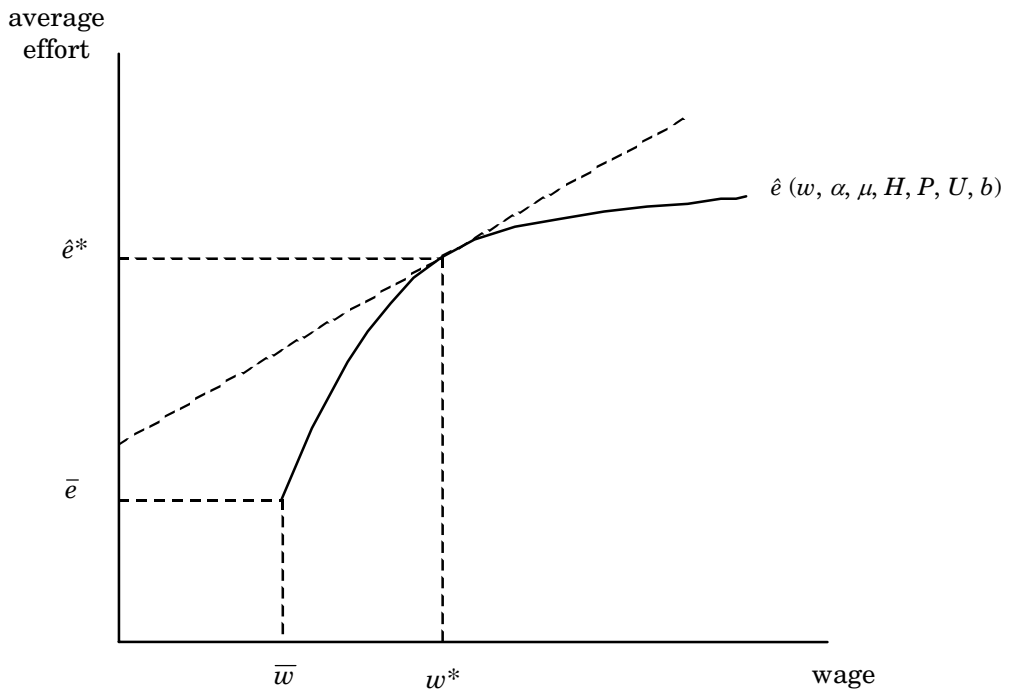
Equation (12') repeats the intuition of Equation (12) in a simpler fashion: without dismissal costs, the last two terms in the denominator of Equation (12) vanish. Equation (13') indicates that the firm sets m so that the induced rise in $\hat{e}$ equals ($\partial \hat{e}/\partial$ w) times μ, the cost of the monitoring input.[26] This solution indicates a point along the firm's expansion path, located at the tangency of an isocost line and the relevant isoquant, reflecting the inputs of wages and monitoring into the production of effort, in this case for a given α. The firm moves along its expansion path by altering labor hours (h) and/or the capital inputs that underlie the output per unit effort (χ); when it does so, Equations (12') and (13') still hold.[27] Equation (14') has an exactly analogous interpretation, inserting the detail that the firm sets α to account for the response of m to α in addition to μ.

Adding replacement costs, as in Equations (12)–(15), complicates the model, but does not alter its fundamental cost-minimizing logic. Accordingly, all three equations adjust the denominators of (12')–(14') to include expected hourly dismissal costs (the last two terms in each denominator; see note 25). Equation (13) adjusts the numerator of Equation (13') to add λFH, the expected cost of dismissal arising from a one-unit increase in m. Similarly, Equation (14) adjusts the numerator of Equation (14') to account for the impact of an increase in α on expected dismissal costs, via its impact on monitoring.[28] Finally, Equation (15) indicates that the firm sets its firing probability to equate $\partial F/\partial \hat{e}$ to the ratio of $\hat{e}$ times λmH—expected dismissal costs generated by a one-unit increase in F—divided by expected hourly costs, without the impact of a change in $\hat{e}$ on F in the denominator.[29]

To summarize, the firm chooses its wage, monitoring input, work standard, and rate of firing in order to maximize the ratio of $\hat{e}$ to total labor costs, accounting for the costs of the other three inputs into the production of effort. In the ensuing discussion, we refer to the full model, with Equations (12)–(15) as model A, and the model without dismissal costs, with Equations (12')–(14') as model B. (The Appendix lists the equations for model B.)

We now proceed to a simplified graphical exposition that illustrates the firm's wage-setting process. Figure 1 shows the worker's utility-maximizing level of effort in response to possible wages, for a given m, α, and F, holding exogenous factors (μ,H,P,U,b) constant. At wage $\overline{w}$, defined as the wage that equates the value of employment with the reservation utility ($v = z$), there is no cost of job loss. Facing no penalty from dismissal, the worker exerts her voluntary level of effort ($\bar{e}$).[30] The firm can elicit more effort by increasing the wage somewhat to w^*, the cost-minimizing wage from Equation (12), shown at the point of tangency in Figure 1. Since $\hat{e}$ is a function of the w (Equation 9), this point simultaneously determines the equilibrium average level of effort, $\hat{e}^*$.

FIGURE 1
The Effort Function



Moving to comparative statics, a change in any of the exogenous terms μ, b, U, or H alters the relationship among w, m, α, F, and $\hat{e}$, shifting and/or rotating the effort function in Figure 1. An increase in b or a reduction in U will lower $\hat{e}$ (Equation (9)). In order to maintain its first order condition, the firm must increase w^* to compensate.[31] Similarly, considering the tangency expressed in Equations (13) or (13'), a rise in μ makes monitoring less attractive (relative to wages) in the production of effort. As the firm substitutes away from m, it will increase w^*.[32] These results are all common to the efficiency wage literature. They are not affected by the introduction of dismissal costs.

Inclusion of replacement costs, as in model A, renders the firm's decision more complicated, reflecting the introduction of worker market power. As indicated above, the firm may decide not to fire a worker found performing insufficiently if replacement

costs are too high. As H increases, the probability of dismissal for a given m, α, and $\hat{e}$ decreases (F falls). The dismissal mechanism, then, becomes a less effective means for generating effort. Consequently, the firm substitutes toward wages in the production of effort.[33] The effort function in Figure 1 shifts to the right, generating a tangency at a higher wage. This induced increase in the w^* reflects the worker's market power, because the worker is costly to replace. This final result, which does not appear in most of the efficiency wage literature, forms the foundation for many of the applications that follow.

INITIAL IMPLICATIONS: EMPLOYERS, WORKERS, AND MARKET POWER

As stated in the literature review, Bowles and Gintis [1990, 177-87] use a labor discipline model without dismissal costs to argue that employers possess unilateral market power over workers. Model B illustrates this argument: the employer uses the threatened sanction of potential job loss to induce effort above the worker's voluntary level.With zero replacement costs, the worker has no similar sanction.[34]

By adding replacement costs, model A renders market power in the employment relationship bilateral, typically increasing the wage. In model A, then, a kind of implicit bargaining takes place between workers and their employers, the outcome of which depends on the relative size of the sanction that each party can impose on the other—on the relative size of the cost of job loss as opposed to hiring and training costs. In most instances, employers have the upper hand since, as a rule, employers face lower costs of exit from individual employment relationships than workers do. An employer may dismiss a worker without leaving the production process itself. By contrast, termination or quitting probably involves at least some period of unemployment for workers. The resulting asymmetry is compounded by (most) workers' dependence on jobs for subsistence. Rarely do employers depend on individual workers for their subsistence.[35] This difference may be further compounded by employers' ability to use monitoring and the implicit threat of job loss to direct the activity of workers. Workers have no comparable mechanism for directing the activity of employers.[36]

There are exceptions. Workers who possess rare general skills, or significant firm-specific skills, may be very costly to dismiss and may face low or virtually no costs to quitting. More generally, workers who are relatively costly to replace will be able to demand and attain higher wages than other workers. At first glance, this may appear to replicate well-known results from human capital theory: more highly trained workers earn more. But a human capital effect would occur in the absence of replacement costs in either a standard human capital model, or an efficiency wage model like model B, provided that worker training enhances the relationship between wages and the value of discretionary effort. Such models, however, would not predict a stronger wage effect from firm-specific human capital than from general human capital. By incorporating replacement costs, model A makes such a prediction: workers who are costly to replace should earn higher wages, *ceteris paribus*, reflecting their implicit power.

The complete model, then, offers a framework for analyzing the relative power of the parties in a bilateral contested exchange employment relationship. Such a framework can

predict relative wage outcomes arising from differences in exogenous characteristics that influence the market power of workers and employers. The ground is set for applications.

SPECIFIC APPLICATIONS

With this foundation, we now turn to implications of the model on the following topics: segmented labor markets and interindustry wage differentials, rising wage inequality, race- and gender-based differentials, and macroeconomic and regional effects. We find that the model presented here is consistent with results found in much of the relevant literature and potentially enhances explanations for these phenomena.

Segmented Labor Markets and Interindustry Differentials

Consider dual labor markets. For simplicity, divide the labor market into two segments: primary and secondary. Following Dickens and Lang [1985], we can distinguish sectors according to returns awarded for education and experience: the primary sector offers noticeable returns to education and experience, typically generating relatively long-term employment relationships for relatively highly skilled workers. The secondary sector, by contrast, offers few if any returns to education and experience, typically generating relatively short-term employment, along with significant competition among a pool of similarly qualified workers.[37] Applying this distinction to model B, we could approximate the segmentation argument made by Bulow and Summers [1986]: the primary sector faces costs to monitoring effort and therefore pays efficiency wages, whereas the secondary labor market clears. Alternately, still using model B, one could differentiate the sectors on the basis of the relative size of (rather than presence of) monitoring costs, allowing the secondary sector to pay low efficiency wages.

Model A, however, generates a second, complementary, and possibly more potent, explanation for higher primary sector wages. If we make replacement costs (H) a positive function of firm-specific human capital and then, following Doeringer and Piore [1985], argue that primary sector employers develop internal labor markets, we would expect primary sector employers to face relatively high H, as their relatively long-term employees acquire firm-specific human capital. Accordingly, model A predicts that primary sector workers will receive higher pay, arising not only from possibly higher monitoring costs (along with possibly higher average levels of human capital), but also from their implicit market power, derived from costs of replacing their firm-specific human capital. If we proceed to assume that replacement costs for secondary sector workers approach zero, reflected in their lack of internal labor market opportunities, we may distinguish between the segments of the labor market as follows: model A applies to the primary sector; model B, probably with relatively low monitoring costs as well, applies to the secondary sector. Secondary sector workers, then, face the unilateral market power of their employers, whereas primary sector workers possess some market power of their own. The wage differential between sectors, then, should exceed any wage discrepancy generated by different average levels of human capital.

Considering dual labor markets from another angle, the complete model could explain why primary sector jobs are more "worker friendly" than secondary sector jobs. Since high replacement costs reduce firing rates and render monitoring less effective, other things equal, we expect primary sector jobs to rely less on monitoring and dismissal threats, *vis-à-vis* wages, to induce effort. In other words, primary sector jobs are more "worker friendly" in the sense that they are more likely to use the "carrot" of high wages than the "stick" of the implicit dismissal threats embodied in high levels of monitoring.[38]

Moving to a disaggregated treatment of industries, this model is consistent with studies that find persistent interindustry wage differentials for workers of similar measured characteristics and even similar occupations, for example, Krueger and Summers [1987, 1988] and Dickens and Katz [1987]. These studies argue that the employment rents found in standard effort models may offer a theoretical explanation: industries with high capital/labor ratios may face high costs of monitoring, inducing higher industry wages, *ceteris paribus*. Model B by itself could achieve this result.

Gibbons and Katz [1992], however, find existing models unsatisfactory for accounting for their observations of industry wage differentials. They state: "Unmeasured-ability models do not motivate findings of strong pairwise correlations between industries that pay high wages and industries that earn large profits, have high capital-to-labour ratios, and are populated by large firms (Katz and Summers, 1989). Efficiency wage models do not motivate the observed high correlation of the industry wage premium across occupations. And rent-sharing models do not motivate the observed similarity of the industry wage structure across countries with very different market systems" [1992, 530].

Model A may be a candidate for resolving this problem. If industries generate industry-specific human capital or unique internal labor markets, replacement costs may differ systematically by industry, generating industry wage differentials. Such differentials could persist across occupations and countries, even in the presence of different national market structures. By combining monitoring and replacement cost arguments, the present model is at least consistent with the observations cited by Gibbons and Katz [1992]: monitoring differences arising from the labor discipline aspect of the model are consistent with the capital/labor ratio-wage correlation, and systematic differences in replacement costs could explain the persistence of industry wage differentials across occupations and nations.

Rising Wage Inequality

It is well documented that wage inequality in the United States has increased dramatically between the mid-1970s and the mid-1990s, and that a rising disparity between the earnings of highly skilled and less-skilled workers constitutes a significant portion of this phenomenon.[39] Several authors [Bound and Johnson, 1992; Berman, Bound, and Grilliches, 1994] have hypothesized that increased skill differentials, in turn, arise from a change in the nature of technical advance in the current period.[40] Contemporary technical change, much of which has involved information technology, has increased the marginal productivity gap between the two types of workers, augmenting the relative demand for highly skilled workers and hence their relative earnings.

While this argument is consistent with many conventional models (and an efficiency wage model like B), model A can take the analysis further. The rising differential may reflect not only an increased disparity in the returns to skill, but also a relative increase in replacement costs for highly skilled workers. In other words, the information-focused nature of technical change in the past 25 years may have exacerbated the market-power differential between high- and low-skilled workers, in addition to affecting their relative demand.

Another group of economists argues that the rising skill differential reflects mounting international competition, particularly from developing nations. Such competition has increased the effective supply of less-skilled labor, exerting downward pressure on wages for relatively unskilled workers.[41] Again, model B could generate this result: an increase in labor supply increases the cost of job loss for relatively unskilled workers. Model A, however, suggests a further avenue of causation: an increase in the effective supply of less-skilled workers may reduce their replacement costs, particularly hiring costs. The rising skill-based wage differential could, then, not only reflect an outward shift of supply against a downward sloping labor demand curve, but also an induced loss of relative market power for affected workers.

Overall, model A appears not only consistent with both the technical change and international competition explanations of the rising skill differential, but is able to offer an additional theoretical rationale based on a rising market power disparity between the two groups.

Race- and Gender-Based Differentials

Model A may offer some insight into the existence of wage discrimination by race or gender in the presence of profit-maximizing behavior.[42] One theory of discrimination argues that occupational segregation limits the access of women and/or minority workers to specific (usually primary sector) jobs and thereby "crowds" them into other (usually secondary sector) jobs [Bergman, 1986; King, 1992; Tomaskovic-Devey, 1993]. If so, white male workers in primary sector jobs face relatively high reemployment probabilities, and therefore lower costs of job loss than their nonwhite or female counterparts, *ceteris paribus*. Model B alone suggests that affected firms would have less market power over white male employees than over others, and therefore may pay white males higher wages than equivalent women or nonwhite workers. Model A adds dimension to this explanation. Women and minority workers who attain access to primary sector jobs would be relatively easy to replace, indicating low replacement costs.[43] If so, women and minority workers would possess less market power than equivalent white male workers, once again offering a possible explanation for wage discrimination that is consistent with profit-maximizing behavior. Note the rent-sharing aspect of this model suggests that competition may not erode the discrimination effect.

Green and McIntosh [1998] present evidence that is consistent with this view of racial discrimination, though their paper focuses on effort, rather than wages, and more on unions than on 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 that is consistent

with model B.[44] While they do not focus their analysis on labor market discrimination *per se*, they find that high levels of "ethnic content" are associated with relatively high effort, *ceteris paribus*. They comment that this result is consistent with efficiency wage theory if one assumes that ethnic workers face a disadvantage in labor markets: facing a higher cost of job loss, ethnic workers will offer more effort to avoid losing their jobs [Green and McIntosh, 1998, 378]. This result is entirely consistent with analysis from both models B and A.

To summarize, a portion of unexplained race- and gender-based wage differentials for otherwise equivalent workers may reflect differences in the relative market power arising from corresponding differences in replacement costs.

Regional and Macroeconomic Changes

Carlin and Soskice [1990, ch. 17] and Blanchard and Katz [1997] use efficiency wage models as a foundation for upward-sloping medium-term aggregate labor supply curves. They proceed to add a labor demand curve (price-determined real wage curve for Carlin and Soskice [1990]) to develop macroeconomic models for determination of the NAIRU (non-accelerating inflation rate of unemployment). The present model would be consistent with the supply relations in either treatment. In model B, falling unemployment reduces the cost of job loss at a given wage, leading to an increase in the efficiency wage and *vice versa*. Model A can again add explanatory dimension if one complicates the model slightly to allow replacement costs to vary inversely with unemployment (let $H = H(U);\ \partial H / \partial U < 0$). In this case, falling unemployment increases workers' market power and wages respond accordingly.[45] Empirical findings of mildly procyclical wages [Card, 1995, Table I] are consistent with these arguments, particularly if one allows downwardly rigid nominal wages to partially offset some of the procyclical wage effect.

Changes in regional, industrial, or occupational unemployment will have similar effects on wages, targeted more narrowly. The "wage curve," whereby regional wages exhibit a negative correlation with regional unemployment rates [Blanchflower and Oswald, 1994] could be developed from model B; indeed Blanchflower and Oswald [1994] use a similar model in a portion of their theoretical discussion. Again, model A can offer additional insight if regional replacement costs vary inversely with regional unemployment ($H_R = H_R(U_R)$). In this case, worker market power may vary inversely with regional unemployment rates, contributing another possible explanation for wage curve effects.

CONCLUSION

This paper develops a model that combines elements of standard effort models with turnover or dismissal cost models. In so doing, it converts the contested exchange implicit in the effort model, whereby employers exhibit unilateral market power, into a contested exchange with bilateral market power. The interaction of the cost of job loss with the replacement costs influences the equilibrium wage. Systematic differences in either of these costs may then contribute to our understanding of a wide variety of phenomena associated with the distribution of labor income and patterns of

wage change. The model replicates results from the literature on segmented labor markets and interindustry wage differentials, but offers additional rationale for their existence based on systematic differences in replacement costs linked to firm- and industry-specific human capital. Similarly, the impacts of recent technological change and rising international competition on skill-based wage differentials may extend beyond demand effects on human capital premiums or effective supply effects on less-skilled wages: both technical change and international competition may reduce the market power of less-skilled workers relative to the more highly skilled. In a similar fashion, race- and gender-based distinctions in replacement costs may contribute to an explanation of the persistence of race- and gender-based wage differentials. Finally, this model is consistent with macroeconomic notions of the NAIRU and with "wage curve" effects of regional unemployment on regional wages, and it adds a market power dimension to their interpretation.

APPENDIX

Model B—Simplified Model, with no Dismissal Costs

Equations (1), (4), (5), (6), (7), and (8) remain as in Model A. Equation (15) does not appear in model B.

(2') $$F(\alpha,\hat{e}) = \Pr(e<\alpha)\big|_{\hat{e}}$$

(3') $$D = D(\alpha,\hat{e};\mu) = M * F = \lambda m(\alpha;\mu) * \Pr(e<\alpha)\big|_{\hat{e}}$$

(9') $$\hat{e} = \hat{e}(D,z,v) = \hat{e}(w,a;\mu,U,b)$$

(10') $$\pi = p*\left[\chi * \hat{e}(w,\alpha;\mu,U,b)\right]h - \left[(w+\mu m)h\right]$$

(11') $$\text{Min}\,\Psi = \left(w+\mu m(\alpha;\mu)\right)\big/\hat{e}$$

(12') $$\partial\hat{e}/\partial w = \hat{e}/(w+\mu m)$$

(13') $$\partial\hat{e}/\partial m = \mu\hat{e}/(w+\mu m) = \mu(\partial\hat{e}/\partial w)$$

(14') $$\partial\hat{e}/\partial\alpha = \left(\mu(\partial m/\partial\alpha)\hat{e}\right)/(w+\mu m) = (\partial\hat{e}/\partial m)(\partial m/\partial\alpha) = \mu(\partial m/\partial\alpha)(\partial\hat{e}/\partial w)$$

NOTES

The author wishes to thank Mark Montgomery and three anonymous referees for helpful comments and suggestions, and Satyendra Patrabansh and Wanlin Liu for research assistance.

1. Implicit contract models [Azariadis, 1975] offer another solution to the principal-agent problem of effort elicitation. The effort model, however stresses costly termination, a key ingredient in the concept of market power developed here.

2. Fair wage models [Akerlof, 1982; Akerlof and Yellen, 1990] resolve the principle-agent problem surrounding effort somewhat differently: because employers understand that groups of workers form norms regarding appropriate effort and that such norms respond to conceptions of fair wages, employers pay efficiency wages to generate motivated effort norms. The formation of group norms, while relevant to the current problem, lies beyond the scope of this paper. For a model that combines fair wages and turnover costs, see Ferguson [forthcoming]. For reviews of the efficiency wage literature, see Akerlof and Yellen [1986] and Weiss [1990].
3. According to Bowles and Gintis [1990], contested exchange occurs whenever the resolution of differences of interest that arise in market transactions requires an enforcement mechanism that is at least partially endogenous to the exchange itself, a condition that becomes necessary when contracts cannot completely stipulate the terms of an exchange. By contrast, a traditional neoclassical labor market assumes *exogenous* enforcement.
4. North [1990, 32] uses a nearly identical notion of power in exchange.
5. In two variations of his model, Hosios [1990, 295] points out that the share distribution of output depends on the relationship between vacancies and unemployment, such that rising vacancies increase labor's share and rising unemployment decreases it.
6. Blanchard and Diamond [1989] derive a Beveridge curve from a matching model that shows wage negotiation after a successful match has been made. Summarizing relevant literature, Blanchard and Katz [1997] argue that matching models imply bargaining power and, moreover, that aggregate employment or unemployment affects such bargaining power. They further argue that efficiency wage models imply bargaining power in a similar fashion.
7. Assumption 2) extends the domain of utility maximizing to include effort. It does not require that workers dislike work, only that workers' utility-maximizing level of effort without monitoring is less than that which would maximize a firm's profits. Additional effort is either not pleasant, or is less desirable at the margin than some other use of energy [Bowles, 1985, 22].
8. Assumptions 1) and 3) are specific exceptions to the standard competitive assumption of perfect information. Stiglitz [1993] refers to effort models as a form of "information economics."
9. Assumption 5) reflects a stylized fact of work: idleness is boring.
10. The monitoring input includes supervisory labor and relevant capital (for example, video cameras).
11. Even if replacement costs are a fixed value, the firm can calculate an expected hourly value over the duration of expected future employment. Later, the idea that H is a positive function of firm-specific human capital and a negative function of unemployment will be considered.
12. Note the second term of the F function = 1 when $H = 0$; whenever $H > 0$, F decreases with a rise in H or a rise in e. A similar F function could arise if we left e constant but made monitoring subject to random mistakes. Gintis and Ishikawa [1987] use such measurement error in their model, making the standard deviation on monitoring a negative function of the monitoring input.
13. Note the role of assumption 5) ($\bar{e} > 0$) in generating this result.
14. In this case, F in Equation (2) would simplify to $F = F(\alpha; e) = \Pr(e < \alpha)$. The dismissal function from Bowles [1985] resembles Equation (3) without F, α, or an explicit normal distribution for e.
15. This equation is similar to that shown in Bowles and Gintis [1990, 213, note #38], with slightly different handling of discounting as well as a specified z function along with α and (H).
16. It can be shown that these partial derivatives have the appropriate signs.
17. This term is equivalent to the cost of job loss in Bowles and Gintis [1990, 178] with discounting handled slightly differently.
18. Gintis and Ishikawa [1987] develop a model in which the probability of dismissal depends on the difference between average group and individual effort ($e_g - e_i$); in a Nash equilibrium, all individuals adopt the group standard. In the present paper, $\hat{e}$ might be interpreted as such a group standard.
19. More formally, one can apply the implicit function rule to Equation (8) to show that $(\partial \hat{e}/\partial w) > 0$. From Equation (8), we let $\psi = (\partial u/\partial \hat{e}) - (\partial D/\partial \hat{e})^{*}\delta(v - z) = 0$. Using the implicit function rule,

$$(\partial \hat{e}/\partial w) = \frac{-(\partial \psi/\partial \hat{e})}{(\partial \psi/\partial w)} = \frac{-(\partial D/\partial \hat{e})\delta(\partial(v-z)/\partial w)\big|_{\hat{e}}}{2\left\{1 + \lambda m\left(\partial^2 F/\partial \hat{e}^2\right)\delta(v-z) + \lambda m(\partial F/\partial e)\delta(\partial(v-z)/\partial \hat{e})\right\}}.$$

Note that the numerator applies for a given $\hat{e}$ since $(\partial \hat{e}/\partial w)$ appears on the left-hand side of the equation. The value is greater than zero since $(\partial(v - z)/\partial \hat{e}) < 0$ (higher effort reduces the cost of job

loss since it reduces the utility of current employment). We can assume that $(\partial^2 F/\partial \hat{e}^2) \geq 0$: diminishing or constant returns of effort with respect to F; recall that $(\partial F/\partial \hat{e}) < 0$. Even if $(\partial^2 F/\partial \hat{e}^2)$ were < 0, $(\partial \hat{e}/\partial w)$ would still be positive as long as $(\partial^2 F/\partial \hat{e}^2) < [1+ \lambda m(\partial F/\partial \hat{e})\delta(\partial(v - z)/\partial \hat{e})]/[(\lambda m \delta(v - z)]$.

20. We assume the condition under note 19 holds. Using the implicit function rule on Equation (8), the equation in the text can be derived.
21. Again, we assume the condition in note 19 and apply the implicit function rule to Equation (8).
22. These results can be shown, using the implicit function rule on Equation (8).
23. As for μ and α, technically, an increase in H has two potentially opposing effects: it reduces the D, but increases v. For an increase in H to reduce $\hat{e}$, the absolute value of the expected effect of the induced decrease in D on the cost of job loss must exceed the induced gain in employment utility: $\left|\left((\partial F/\partial H)\big|_{\hat{e}}\right)\delta(v - z)\right| > \left\{\left(\partial(v - z)/\partial F\right) * \left(\partial F/\partial H\right)\right\}\big|_{\hat{e}}$. This simplifies to: $\delta(v - z) > \left(\partial(v - z)/\partial F\right)\big|_{\hat{e}}$.
24. LaJeunesse [1999] suggests that effort and hours are not independent; reducing the hours of the workweek would increase effort. That argument, however, will not be pursued here.
25. The third term in the denominator of Equation (12) shows the expected hourly costs of dismissal holding effort constant. The last term represents the interaction effect of the increase in effort, induced by the wage increase, on the probability of dismissal from Equation (3). As effort increases, F falls, causing expected dismissal costs to fall by the amount shown.
26. Alternately, $(\partial \hat{e}/\partial m)/\mu = \hat{e}/(w + \mu m) = (\partial \hat{e}/\partial w)$, using Equation (12'). Put slightly differently, the ratio of monitoring costs to wage costs $(\mu/1 = \mu)$ is equal to the marginal rate of substitution between wages and monitoring in the production of effort $(\partial \hat{e}/\partial m)/(\partial \hat{e}/\partial m)$.
27. Bowles [1985, 24] makes these points for a similar model without α.
28. The first new term in the numerator, $\lambda(\partial m/\partial \alpha)FH$, shows the response of monitoring frequency (M) to a change in α times expected dismissal costs (FH), holding $\hat{e}$ constant. The second new term, $\left(\lambda m H(\partial F/\partial \alpha)\big|_{\hat{e}}\right)$, indicates an equivalent effect of α on F.
29. The denominator of Equation (15) is equivalent to the denominators of Equations (12)–(14) with $(\partial F/\partial \hat{e}) = 0$, since its inverse appears on the left-hand side of Equation (15).
30. $\overline{w}$ is not a Walrasian market clearing wage, since assumptions 1-3 indicate a non-Walrasian world [Bowles and Gintis, 213, note 36]. It is impossible, therefore, to tell whether w^* is above or below a hypothetical Walrasian market-clearing wage. However, $\overline{w}$ is a reservation wage in the sense that the worker will refuse employment at any wage below $\overline{w}$.
31. Concerning an exogenous rise in unemployment, along a given isoquant, where $d\hat{e} = 0$, $(\partial w/\partial U) = -(\partial \hat{e}/\partial U) < 0$. An analogous argument applies to b. If a rise in b or a drop in U should increase H, as could be argued, complicating the model slightly, the attendant rise in w^* from model A exceeds that from model B.
32. From Equation (13'), as μ increases, the ratio $(\partial \hat{e}/\partial m)/\mu$ falls, necessitating a drop in $\partial \hat{e}/\partial w$. The wage increases since $\partial^2 \hat{e}/\partial w^2 < 0$. Equation (13) yields the same result since λFH is independent of μ. Further, if we were to assume a positive correlation between H and μ, a rise in μ would induce a somewhat larger wage increase in model A than in model B.
33. Using reasoning analogous to note 32 with Equation (15), a rise in H causes the ratio $(\partial \hat{e}/\partial F)/H$ to decrease, requiring a drop in the ratio $(\lambda m \hat{e})/(w+\mu m+\lambda m FH)$. The latter ratio is equal to $\lambda m(\partial \hat{e}/\partial w)$, with $(\partial F/\partial \hat{e}) = 0$ (as in the denominator of Equation (15)). The ensuing drop in $(\partial \hat{e}/\partial w)$—see note 32—requires a rise in w^* since $(\partial^2 \hat{e}/\partial w^2) < 0$.
34. The worker's effort constrains the firm's choice of a profit-maximizing wage, but the effort function is a passive response function, analogous to a Stakelberg follower [Bowles and Gintis, 1990]. The wage-effort relationship represents the worker's autonomy, not power, since the worker is not in a position to impose costs on the firm [Bowles and Gintis, 1990, 183]. The firm may dismiss and replace the worker at no cost to the firm.
35. Skillman [1991] applies this argument to exit options in a Rubinstein bargaining model. Bacharach and Lawler [1981] base their theory of bargaining power on dependence: the asymmetric dependence of individual workers on their employer places them in an inferior bargaining position.
36. These differences may be expressed in terms of the exit/voice categories developed by Hirschman [1970]. Dismissal and quitting are forms of exit for firms and workers, respectively. Monitoring is a form of employer voice.

37. Dickens and Lang [1985] empirically distinguish the sectors on the basis of returns to education and experience. Albrecht and Vroman [1992] use an efficiency wage/search model endogenously to generate a dual labor market with a separating equilibrium between sectors. Their model indicates that the primary sector rations jobs and pays a higher wage and requires more effort than the secondary sector. The present notion of segmentation is consistent with both of these concepts of a dual labor market (indeed they are consistent with each other), but it adds the notion that replacement costs, arising from firm-specific human capital, may play a role in generating wage differences arising from segmentation.
38. To offer a stark example, consider the relative importance of implicit dismissal threats for inducing effort from tenured faculty members as opposed to fast food workers. The argument that the relative importance of wages *vis-à-vis* monitoring is higher for the primary sector does not imply an absence of monitoring for the primary sector; with a large cost of job loss, evident monitoring is less important for inducing effort.
39. See Levy and Murnane [1992] for an excellent review of this literature. See also Danziger and Gottschalk [1993] and Gottschalk [1997].
40. Juhn, Murphy and Topel [1991] make a similar, largely compatible argument that increased relative demand for high-skilled labor has led to an increase in the NAIRU (non-accelerating inflation rate of unemployment)—via increasing unemployment for the less-skilled—and rising skill-based wage differentials. They do not specifically link the demand differentials to technological change, however.
41. See, for example, Freeman [1995], Wood [1995], and Feenstra and Hanson [1996].
42. Another critical arena for discrimination involves promotion practices. Promotion issues are not well addressed by the static models presented here, though the notions of market power presented here should be relevant.
43. Forces external to the firm may generate systematic race- or gender-based differences in H. Workers remain identical with respect to productivity and preferences.
44. Green and McIntosh's [1998] notion of a cost of job loss is also consistent with Model A, though they do not specifically consider costs of dismissal.
45. Note that in model B, where employers enjoy unilateral power, falling unemployment reduces the potency of dismissal threats at a given wage: the employer's power varies directly with U. Model A adds the dimension that the worker's power varies inversely with U, compounding the impact of changes in U.

REFERENCES

Akerlof, G. A. Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics*, November 1982, 543-69.

Akerlof, G. A. and Yellen, J. The Fair Wage-Effort Hypothesis and Unemployment. *Quarterly Journal of Economics*, May 1990, 255-83.

_____, eds. *Efficiency Wage Models of the Labor Market*. Cambridge, U.K.: Cambridge University Press, 1986.

Albrecht, J. W. and Vroman, S. B. Dual Labor Markets, Efficiency Wages, and Search. *Journal of Labor Economics,* October 1992, 438-61.

Azariadis, C. Implicit Contracts and Unemployment Equilibria. *Journal of Political Economy,* December 1975, 1183-1202.

Bacharach, S. B. and Lawler, E. J. *Bargaining: Power, Tactics and Outcomes*. San Francisco: Jossey Bass Publishers, 1981.

Bergmann, B. *The Economic Emergence of Women*. New York: Basic Books, 1986.

Berman, E., Bound, J., and Grilliches, Z. Changes in the Demand for Skilled Labor within U.S. Manufacturing Industries. *Quarterly Journal of Economics*, May 1994, 367-98.

Blanchard, O. and Diamond, P. The Beveridge Curve. *Brookings Papers on Economic Activity*, 1989 (1), 1-60.

Blanchard, O. and Katz, L. F. What We Know and Do Not Know About the Natural Rate of Unemployment. *Journal of Economic Perspectives*, Winter 1997, 51-72.

Blanchflower, D. G. and Oswald, A. *The Wage Curve.* Cambridge, Mass.: MIT Press, 1994.

Bound, J. and Johnson, G. Changes in the Structure of Wages in the 1980's: An Evaluation of Alternative Explanations. *American Economic Review*, June 1992, 371-92.

Bowles, S. The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models. *American Economic Review*, Spetember 1985, 16-36.

Bowles, S. and Gintis, H. Contested Exchange: New Micro-foundations for the Political Economy of Capitalism. *Politics and Society*, June 1990, 165-222.

Bulow, J. I. and Summers, L. H. A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination and Keynesian Unemployment. *Journal of Labor Economics*, July 1986, 376-414.

Cahill, M. B. Exploring the Interaction between Efficiency Wages and Labor Market Frictions. *Quarterly Review of Economics and Finance*, Spring 2000, 121-37.

Card, D. The Wage Curve: A Review. *Journal of Economic Literature*, June 1995, 785-99.

Carlin, W. and Soskice, D. *Macroeconomics and the Wage Bargain.* Oxford: Oxford University Press, 1990.

Danziger, S. and Gottschalk, P., eds. *Uneven Tides: Rising Inequality in America.* New York: Russell Sage Foundation, 1993.

Dickens, W. T. and Katz, L. F. Inter-Industry Wage Differences and Industry Characteristics, in *Unemployment and the Structure of Labor Markets*, edited by K. Lang and J. S. Leonard. New York: Basil Blackwell, 1987, 48-89.

Dickens, W. T. and Lang, K. Testing Dual Labor Market Theory: A Reconsideration of the Evidence. Working Paper 1670, National Bureau of Economic Research, 1985.

Doeringer, P. B. and Piore, M. J. *Internal Labor Markets and Manpower Analysis.* Armonk, N.Y.: M.E. Sharpe, Inc., 1985.

Feenstra, R. C. and Hanson, G. H. Globalization, Outsourcing, and Wage Inequality. *American Economic Review, Papers and Proceedings*, May 1996, 240-45.

Ferguson, W. D. Fair Wages, Worker Motivation and Implicit Bargaining Power in Segmented Labor Markets. *Journal of Theoretical and Institutional Economics*, forthcoming.

Freeman, R. B. Are Your Wages Set in Beijing? *Journal of Economic Perspectives*, Summer 1995, 15-32.

Gibbons, R. and Katz, L. Does Unmeasured Ability Explain Inter-Industry Wage Differentials? *The Review of Economic Studies*, July 1992, 515-35.

Gintis, H. and Ishikawa, T. Wages, Work Intensity, and Unemployment. *Journal of Japanese and International Economies* 1, 1987, 195-228.

Gottschalk, P. Inequality, Income Growth, and Mobility: The Basic Facts. *Journal of Economic Perspectives*, Spring 1997, 21-40.

Green, F. and McIntosh, S. Union Power, the Cost of Job Loss, and Workers' Effort. *Industrial and Labor Relations Review*, July 1998, 363-83.

Hirschman, A. O. *Exit, Voice and Loyalty*. Cambridge, Mass.: Harvard University Press, 1970.

Hosios, A. J. On the Efficiency of Matching and Related Models of Search and Unemployment. *The Review of Economic Studies*, April 1990, 279-98.

Juhn, J., Murphy, K. M., and Topel, R. H. Why has the Natural Rate of Unemployment Increased over Time? *Brookings Papers on Economic Activity*, (2) 1991, 75-126.

King, M. C. Occupational Segregation by Race and Sex 1940-88. *Monthly Labor Review*, April 1992, 30-37.

Krueger, A. B. and Summers, L. H. Efficiency Wages and the Inter-Industry Wage Structure. *Econometrica*, March 1988, 259-93.

_____. Reflections on the Inter-Industry Wage Structure, in *Unemployment and the Structure of Labor Markets*, edited by K. Lang and J. S. Leonard. New York: Basil Blackwell, 1987, 17-47.

LaJeunesse, R. Toward an Efficiency Week. *Challenge*, January-February 1999, 92-109.

Levy, F. and Murnane, R. J. U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations. *Journal of Economic Literature*, September 1992, 1333-81.

Lindbeck, A. and Snower, D. J. *The Insider-Outsider Theory of Employment and Unemployment.* Cambridge, Mass.: MIT Press, 1988.

North, D. *Institutions, Institutional Change and Economic Performance, The Political Economy of Institutions and Decisions.* Cambridge, U.K.: Cambridge University Press, 1990.

Salop, S. C. A Model of the Natural Rate of Unemployment. *American Economic Review,* January 1979, 117-25.

Shapiro, C. and Stiglitz, J. H. Equilibrium Unemployment as a Worker Discipline Device. *American Economic Review*, June 1984, 433-44.

Skillman, G. Efficiency vs. Control: A Strategic Bargaining Analysis of Capitalist Production. *Review of Radical Political Economics*, Winter 1991, 12-21.

Stiglitz, J. Post Walrasian and Post Marxian Economics. *Journal of Economic Perspectives*, Winter 1993, 109-14.

Tomaskovic-Devey, D. *Gender & Racial Inequality at Work*. Ithaca, N.Y.: ILR Press, 1993.

Weiss, A. *Efficiency Wages: Models of Unemployment, Layoffs and Wage Dispersion*. Princeton, N.J.: Princeton University Press, 1990.

Wood, A. How Trade Hurt Unskilled Workers. *Journal of Economic Perspectives,* Summer 1995, 57-80.