Monitoring and Pay

Magnus Allgulin* Tore Ellingsen**

SSE/EFI Working Paper Series in Economics and Finance
No. 245

November, 1999

Abstract

The shirking model of efficiency wages has been thought to imply that monitoring and pay are substitute instruments for motivating workers. We demonstrate that this result hinges critically on restrictive assumptions regarding workers' choice of effort – for example that there are only two possible effort levels. Under more reasonable assumptions, monitoring and pay are complementary instruments. Another result is that there is a non–monotonic relationship between the wage level and the workers' rents. Finally, much of the empirical literature on the monitoring–pay relationship is shown to be seriously misguided.

JEL CLASSIFICATION: J31, J41.
KEYWORDS: Monitoring, efficiency wages, incentive pay.

*Address: Department of Economics, Stockholm School of Economics, Box 6501, S—113 83 Stockholm, Sweden. Email: magnus.allgulin@hhs.se.

**Address: Department of Economics, Stockholm School of Economics, Box 6501, S—113 83 Stockholm, Sweden. Email: gte@hhs.se.

We are grateful to Guido Friebel, Bengt Holmström, Eva Meyerson, Sten Nyberg, Canice Prendergast, an anonymous referee and seminar participants at Stockholm School of Economics and at the conference "Wage Levels and Wage Dispersion" in Sigtuna 1997 for helpful comments and discussion. Financial support from the Swedish Council for Research in the Humanities and Social Sciences, the Stockholm School of Economics and the Royal Swedish Academy of Sciences is gratefully acknowledged.

1 Introduction

Conventional wisdom says that monitoring and pay are substitute instruments for motivating workers: workers who are poorly monitored have to be well paid in order not to shirk, and conversely, workers who are closely monitored do not have to be particularly well paid. This version of the efficiency wage idea has a long and distinguished history. It is also the cornerstone of Bulow and Summers' (1986) theory of dual labor markets. In short, their theory says that there is a primary sector in which monitoring is difficult—hence working conditions are good and pay is high; and there is a secondary sector in which monitoring is easy—hence working conditions are poor and pay is low.

In this paper we shall argue that the accuracy of monitoring and the level of pay are negatively related only under very restrictive assumptions. Notably, the above argument is correct when workers only have a choice between two effort levels (working and shirking), but incorrect when worker effort can be adjusted in a continuous fashion. Similarly, the argument is correct when the desirable level of worker effort is given exogenously, but incorrect when the desirable level of worker effort emerges as part of the solution to the firm's profit maximization problem.

The point is that monitoring costs affect not only the choice between the carrot (pay) and the stick (monitoring) in implementing a given level of effort, but also the level of effort itself. In general, there is a scale effect in addition to the substitution effect.[1] For example, as monitoring becomes more costly, there will be a shift from monitoring toward wage incentives, just as Bulow and Summers claim, but there will also be a reduction in the desired effort level. Without further analysis, the total effect on the wage is thus ambiguous. How far can further analysis take us?

Briefly stated, our main contribution is to build a simple shirking model with a continuum of effort levels and to characterize the profit maximizing levels of monitoring and pay. Interestingly, complementarity holds regardless of which of the model's parameters is varied. The main finding is therefore that, under weak assumptions, monitoring and pay are complementary instruments. A parameter change which causes an increase in the accuracy of monitoring also causes an increase in the level of pay and vice versa. Our model also demonstrates that the discrete shirking model can not be viewed as a limiting case of a continuous model. We conclude that the "simplifying" assumption of discrete effort levels is neither innocent nor justifiable.

[1] The possibility that scale effects could outweigh substitution effects has been mentioned earlier by Rebitzer (1993,1995). However, as far as we know, the issue has not been analyzed in any detail.

Earlier theoretical literature has taken it for granted that workers are better off in firms where it is difficult to monitor and in firms which pay high wages. Neither of these propositions is true here. Workers are best off in intermediate cases. In firms where monitoring is either very difficult or very easy, workers are paid close to their reservation wage. The difference is that when monitoring is easy, workers are required to work harder and therefore receive a higher wage. Indeed, the very highest wages will be paid by firms which monitor very accurately, and hence do not need to pay a wage premium.

The paper also identifies three problems associated with empirical implementation of efficiency wage theory. First, the accuracy of monitoring is very hard to measure. Most studies proxy the accuracy of monitoring by the intensity of supervision, as measured by the number of supervisors per worker or the frequency of supervision.[2] This is illegitimate. The number of supervisors may be high precisely because monitoring is difficult. Existing studies have essentially investigated the relationship between monitoring costs and pay. Unfortunately, this relationship is more complicated than the relationship between the monitoring level and pay. Here is a second problem: if there is sufficient variation in the underlying (unmeasured) parameters, there will be two possible wage levels for each level of monitoring costs. Thus, unless the effort level is properly controlled for, the empirical literature's regressions of wage premiums on monitoring costs are unjustified. But all is not lost. A more constructive contribution of our work is to show that it is still legitimate to run a regression of monitoring costs on wage premiums. In other words, existing data sets can still be of value in testing the theory if only regressions are run in reverse! A final obstacle to empirical testing is that entry costs create a sample selection bias. With high entry costs, the unobserved parameters of the model tend to be negatively correlated, because firms with favorable demand characteristics can survive even with poor cost characteristics.

Our basic model can be extended in a number of directions. The extensions we pursue involve imposing plausible restrictions on the feasible set of monitoring levels. These extensions enable us to link the theoretical model quite closely to the empirical work of Groshen and Krueger (1990) on supervision and pay in hospitals and to that of Krueger (1991) on franchising.

The paper is organized as follows. Section 2 sets up the model and studies how variations in monitoring costs affect the levels of monitoring and pay. Here, we also investigate the relationship between the continuous model and the discrete model. Welfare implications are discussed in Section 3. In Sec-

[2] See e.g. Leonard (1987), Gordon (1990,1994), Kruse (1992), Neal (1993).

tion 4, we allow for other sources of variation than the cost of monitoring. Section 5 discusses the model's implications and how they should be tested. We also discuss an issue which is important in some applications, namely, discrete monitoring technologies (organizational forms). Section 6 concludes.

2 Model

The set–up is a simple efficiency wage model in the spirit of Shapiro and Stiglitz (1984). A risk neutral principal employs a single risk neutral agent. The agent can exert effort $e \in \mathbb{R}_+$ which affects the principal's benefit, $\beta B(e)$, at some cost to the agent $\gamma C(e)$. By assumption, the parameters β and γ are both positive. The principal motivates the agent through a compensation contract $w(e)$, where w is the wage that the agent receives if the principal observes e.

The principal can observe and verify the agent's effort with probability p. This probability is affected by the principal's choice of monitoring technology. In order to attain probability p of observing the agent's effort, the principal has to pay $\mu M(p)$, where $\mu > 0$. We make the following assumption regarding functional forms.

Assumption 1 (i) $B'(e) > 0, B''(e) \leq 0$, (ii) $C(0) = 0, C'(e) > 0, C''(e) > 0$, (iii) $M'(p) > 0$.

The ex post utility of the principal can now be written

$$U = \beta B(e) - w - \mu M(p), \tag{1}$$

and the ex post utility of the agent is

$$V = w - \gamma C(e). \tag{2}$$

A key assumption of the efficiency wage model is that there is a lower limit, $w_0 \in \mathbb{R}$, to the payment. The limit may be due either to legal rules or to a wealth constraint.

Since effort is not always observed, the compensation contract also needs to specify some payment $\bar{w} \in \mathbb{R}$ that the agent is to receive in this case. The agent is assumed to maximize expected utility,

$$\mathrm{E}[V] = pw(e) + (1-p)\bar{w} - \gamma C(e). \tag{3}$$

2.1 Discrete effort levels

For a moment assume that the effort level is restricted to take only the values $e = 0$ and $e = \bar{e}$. This is the case in earlier versions of the efficiency wage model. Then the following incentive constraint must be satisfied

$$pw(\bar{e}) + (1-p)\bar{w} - \gamma C(\bar{e}) \geq pw(0) + (1-p)\bar{w}.$$

The principal will optimally punish the agent as much as possible if he is detected cheating, thus $w(0) = w_0$. We impose the standard assumption that an indifferent agent takes the action that the principal favors. The principal can then lower the wage down to the level where the incentive compatibility constraint becomes an equality and the incentive constraint can be simplified to

$$p(w - w_0) = \gamma C(\bar{e}), \tag{4}$$

where $w(\bar{e})$ is denoted by w. This constraint tells us that, as long as the right hand side is constant, an increase in the equilibrium value of monitoring must be followed by a decrease in the equilibrium value of the wage and vice versa. In Figure 1 we have plotted the points of w and p which are just sufficient to implement the effort level $\bar{e}$. After a decrease in μ the principal will move from the initial equilibrium A along the iso-effort curve to point B, substituting wage for monitoring accuracy in order to minimize costs. That is, monitoring and pay are substitute instruments.

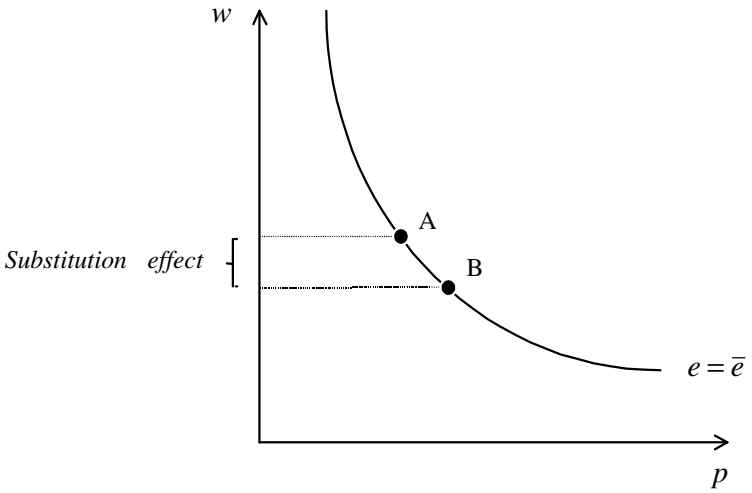

Figure 1

2.2 Continuous effort levels

As we shall see, the above result is an artefact of the very strong restriction that the effort level can only take two possible values. Under weak assumptions, there is a scale effect that dominates the substitution effect, i.e., p and w are complementary instruments with respect to changes in μ. Figure 2 illustrates the result. There, we have plotted two iso-effort curves for different levels of effort. After a decrease in μ, cost minimization implies a move from the initial equilibrium A, along the iso-effort curve to point B. However, at this point the marginal revenue is not equated to the marginal cost of effort, which is required for profit maximization. In order to maximize profits, the principal has to increase the effort level, ending up at point C. The substitution of pay for monitoring accuracy is dominated by the principal's complementary use of monitoring accuracy and pay to increase the implemented level of effort.

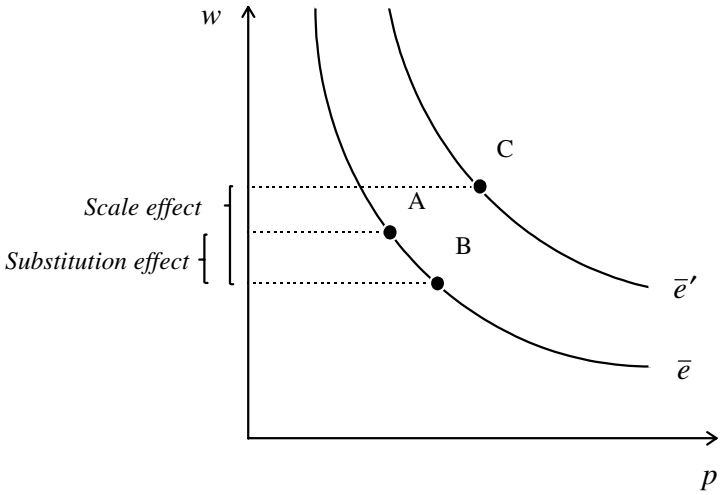

Figure 2

Let us now prove this claim more formally. When the principal is free to induce the agent to take any level of effort $\hat{e} \in \mathbb{R}_+$, the following incentive compatibility constraint must be satisfied for all e:

$$pw(\hat{e}) + (1-p)\bar{w} - \gamma C(\hat{e}) \geq pw(e) + (1-p)\bar{w} - \gamma C(e).$$

We see that any incentive compatible contract that implements $\hat{e}$ can, without loss to the principal, be replicated by a step function of the form $w(e) = w_0$ for $e < \hat{e}$ and $w(e) = w$ for $e \geq \hat{e}$. I.e., the principal sets an effort target,

$\hat{e}$. The agent gets w if he meets or exceeds the target and the minimum payment w_0 otherwise.[3] This contract is illustrated in Figure 3.

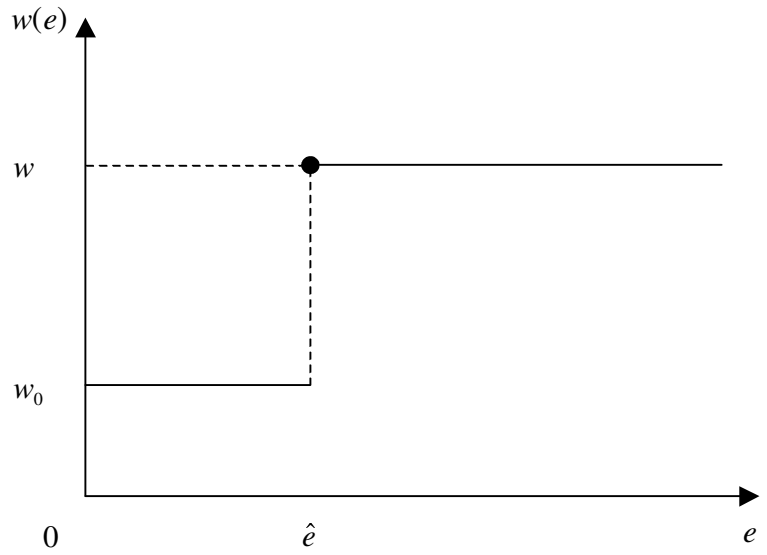

Figure 3

Given this wage contract, if an agent ever wants to deviate, he will deviate to $e = 0$. Thus, the incentive compatibility constraint becomes

$$p(w - w_0) \geq \gamma C(\hat{e}). \tag{5}$$

Finally, the assumption that an indifferent agent takes the action favored by the principal enables the principal to lower the wage down to the level where the incentive compatibility constraint becomes an equality. Inverting this equality, we obtain an expression for the actual effort which the agent will exert,

$$e(p, w) := C^{-1}((w - w_0)p/\gamma). \tag{6}$$

Note that $\bar{w}$, the wage in the case that the effort level is not observed, is irrelevant for the agent's incentives. It is common to assume that $\bar{w} = w(\hat{e})$, and we follow this practice. One reason why the assumption is plausible is that if $\bar{w} < w(\hat{e})$, the principal would have had an incentive not to monitor (or to falsely claim that he has not monitored).

Before we turn to the detailed analysis, it is useful to define

$$r(e) := \frac{\beta B'(e)}{\gamma C'(e)},$$

[3] As shown by Demougin and Fluet (1997), the optimality of effort targets is quite general. It is not an artefact of the simple monitoring technology.

i.e. the ratio of the marginal benefit to the marginal cost. Needless to say, this ratio would be equal to one with perfect monitoring.

The principal's problem is to find a probability p and a wage w to maximize

$$U(p, w) = \beta B(e(p, w)) - w - \mu M(p) \tag{7}$$

subject to the constraints $w \geq w_0$ and $p \in [0, 1]$. This is a straightforward maximization problem in two variables. Let a solution to this problem be denoted (p^*, w^*), and let $e^* := e(p^*, w^*)$ denote the associated effort level. The first–order conditions for the solution can then be written

$$p^* r(e^*) - 1 \leq 0, \tag{8}$$

with equality if $w^* > w_0$, and

$$(w^* - w_0) r(e^*) - \mu M'(p^*) \geq 0, \tag{9}$$

with equality if $p^* < 1$.

Equation (8) tells us that the marginal benefit from increased effort will be larger than the marginal cost whenever the principal chooses to monitor imperfectly. The reason for the distortion away from the socially optimal effort level is that the agent must be paid a rent in order not to shirk. Also, equation (8) confirms that there must be a positive level of monitoring in order to induce any effort. (If $p^* = 0$, then $w^* = w_0$.)

Let U_{ij} denote the second derivative of U and let U^*_{ij} denote the second derivative of U evaluated at a solution (p^*, w^*). The second–order conditions are then $U^*_{ww} < 0, U^*_{pp} < 0$ and $U^*_{pp} U^*_{ww} - (U^*_{pw})^2 > 0$. To state the conditions in full, note that

$$U_{ww} = \frac{\beta (p)^2 h(e)}{\gamma^2}, \tag{10}$$

$$U_{pp} = \frac{\beta (w - w_0)^2 h(e)}{\gamma^2} - \mu M''(p), \tag{11}$$

$$U_{pw} = \frac{\beta p (w - w_0) h(e)}{\gamma^2} + r(e), \tag{12}$$

where

$$h(e) := \frac{B''(e) C'(e) - C'''(e) B'(e)}{(C'(e))^3}.$$

2.3 Variation in monitoring costs

Our main objective is to characterize how the wage level and the accuracy of monitoring vary with the parameter μ. Given that the solution is interior, this is a standard comparative static exercise.

The effect of a change in the cost of monitoring, μ, is found by differentiating the two first–order conditions (8) and (9) to get the two equations

$$\begin{aligned} U^*_{ww}dw^* + U^*_{pw}dp^* &= 0 \\ U^*_{pw}dw^* + U^*_{pp}dp^* &= M'(p^*)d\mu. \end{aligned}$$

Using Cramer's rule, we then have

$$\frac{dw^*}{d\mu} = \frac{-U^*_{pw}M'(p^*)}{U^*_{ww}U^*_{pp} - (U^*_{pw})^2} \tag{13}$$

and

$$\frac{dp^*}{d\mu} = \frac{U^*_{ww}M'(p^*)}{U^*_{ww}U^*_{pp} - (U^*_{pw})^2}. \tag{14}$$

As one would expect, an increase in μ always lowers the level of monitoring. In the other case, the sign of the cross–derivative U^*_{pw} plays a crucial role. Following common usage, we say that monitoring and pay are Edgeworth complements (substitutes) if U^*_{pw} is positive (negative). Thus, our first result can be stated as follows.

Lemma 1 *If $U_{pw} > 0$, monitoring and pay are complementary instruments.*

We are left with the question: How realistic is it that monitoring and pay are Edgeworth complements?

We claim that monitoring and pay are always Edgeworth complements if the principal's benefit function, $B(e)$, and the agent's cost of effort function, $C(e)$, are both represented by any kind of power function satisfying our initial assumptions. To see this, just let $B(e) = b_0e^{b_1}$ and $C(e) = c_0e^{c_1}$. Now, it follows from (6) and (7) that

$$U = \beta b_0[(w - w_0)p/\gamma c_0]^{b_1/c_1} - w - \mu M(p),$$

so

$$U_{pw} = \frac{\beta b_0 b_1^2}{\gamma c_0 c_1^2}\left[\frac{(w - w_0)p}{\gamma c_0}\right]^{\frac{b_1 - c_1}{c_1}},$$

which is clearly positive if both b_0 and c_0 are positive.

In the commonly studied special case that the principal's benefit function is linear, a more general statement can be made, namely that monitoring and pay are Edgeworth complements, and hence complementary instruments, if and only if the relative growth of cost of effort is decreasing in the effort level.

To prove this formally, let $B(e) = b_0 e$. It follows that $B'(e) = b_0$ and that $B''(e) = 0$. Now monitoring and pay are Edgeworth complements if and only if (12) is positive in equilibrium, i.e. if and only if

$$U^*_{pw} = -\frac{\beta b_0 p^*(w^* - w_0)C''(e^*)}{\gamma^2(C'(e^*))^3} + \frac{\beta b_0}{\gamma C'(e^*)} > 0.$$

Use the IC-constraint, (6), to rewrite the above condition to

$$U^*_{pw} = \frac{\beta}{\gamma}\frac{(C'(e^*))^2 - C(e^*)C''(e^*)}{(C'(e^*))^3} > 0. \tag{15}$$

That is, $U^*_{pw} > 0$ if and only if

$$\frac{\partial \frac{C'(e^*)}{C(e^*)}}{\partial e^*} < 0. \tag{16}$$

To sum up, given that the parameter μ is the source of variation, we have found two differerent sufficent conditions for monitoring and pay to be complementary instruments.

Proposition 1 *Sufficient conditions for monitoring and pay to be complementary instruments are (each effective individually):*

1) The principal's benefit function, $B(e)$, and the agent's cost of effort function, $C(e)$, can both be represented by power functions.

2) The principal's benefit function, $B(e)$, is linear and the relative growth of cost of effort is decreasing in the effort level.

As one would expect, these conditions are sufficient for complementarity if the source of variation is β or γ as well, but we postpone a formal proof of this result until Section 4.

We think that both the conditions are quite plausible. Condition 1) is perhaps particularly useful because power functions are conventionally used in principal–agent models; e.g., the linear–quadratic model satisfies this assumption. Condition 2), which also encompasses the linear–quadratic model, is of particular relevance if the principal is a price taker in the output market. If we measure effort in terms of output, the benefit function is obviously linear. The requirement on $C(\cdot)$ is then satisfied by a large set of functions. For

example, it is satisfied by any polynomial function with positive coefficients. We conclude that monitoring and pay are complementary instruments under mild assumptions, in stark contrast with previous claims that are based on a model with discrete effort levels.

A common defense of discrete models is that they can be thought of as a limiting case of a continuous model. For example, if the function $C(\cdot)$ is strongly convex over some range, it would be as if only two effort levels were relevant. We end the section by demonstrating that this argument should be used cautiously. Not all discrete models can be given this justification. In particular, Shapiro and Stiglitz' model can not be given this interpretation. Let the benefit function and the cost of effort function be represented by power functions as above, and let the parameter c_1 go to infinity. It follows that $C(e > 1) = \infty$, $C(e = 1) = c_0$ and $C(e < 1) = 0$; thus the only two relevant levels of effort are $e = 1$ and $e = 0$. To find the sign of U_{pw}, just let $c_1 \to \infty$ in the expression (15), whereupon it is seen that $U_{pw} \to 0_+$, i.e. converges to zero *from above*. An extremely convex $C(\cdot)$ plainly implies that the cost of implementing the high effort level is approximately the same as the cost of implementing the low effort level. It follows that the elasticity of substitution between monitoring and pay must be approximately zero. Perhaps this is most easily seen with reference to our treatment of the discrete case in Section 2.1. In order to implement the high effort $\bar{e}$, the expected added payment to working, $p(w - w_0)$ had to be at least as large as the cost of the added effort $\gamma C(\bar{e})$. When the latter term goes to zero, the iso–effort curve described by equation (4) is squeezed towards the origin of the (p, w)–space. Thus U_{pw} remains positive, and monitoring and pay are complementary instruments. The discrete models of Shapiro and Stiglitz (1984) and Bulow and Summers (1986) maintain the assumption that the cost to the principal of implementing high effort is "significantly" higher than the cost of implementing low effort. Hence these models can *not* be interpreted as approximations of a continuous model.

3 Welfare implications

Most work on dual labor markets has taken for granted that workers are better off in jobs that are poorly monitored. Here, this is not true at all. From (8) and (9), we have that

$$w^* = w_0 + p^* \mu M'(p^*), \tag{17}$$

and from (5) we know that

$$\gamma C(e(p^*, w^*)) = p(w^* - w_0). \tag{18}$$

The agent's utility can hence be written

$$\begin{aligned} V &= w^* - \gamma C(e(p^*, w^*)) \\ &= w_0 + p^*(1-p^*)\mu M'(p^*). \end{aligned} \tag{19}$$

Thus, for $p^* = 0$ and for $p^* = 1$ the agent's utility is at its minimum level, w_0. For all other levels of monitoring, the agent earns rents. The point is very simple: if the principal cannot monitor at all, the agent will always shirk, and there is no reason to pay an efficiency wage. Conversely, if the principal can monitor perfectly, any effort level can be implemented as long as the agent is paid the corresponding reservation wage.

By implication, the agent's utility is a non–monotonic function of w^*. The wage is increasing in the level of monitoring, and hence utility is at its lowest for very low and very high wages. The best paid workers are required to put in a lot of effort. Hence, the true rents are largest in the intermediate wage range.

If the model is an appropriate description of reality, it means that workers need not be worse off when monitoring technologies improve, at least as long as monitoring remains reasonably imperfect. This may be one reason why the mechanization of production, contrary to the claims of Braverman (1974) and other Marxian scholars, does not seem to have implied increased exploitation of workers.[4]

4 Other sources of variation

Another obstacle to empirical testing of the model is that μ is not the only parameter of the model. What if γ changes instead? Even in a model with only two effort levels we can show that monitoring and pay are complements in this case. Figure 4 illustrates the outcome. An increase in the cost of effort, represented by a move from γ to γ', means that the agent requires a stronger incentive to maintain the effort level $\bar{e}$. This can be accomplished by higher wages or by more monitoring. It turns out that the principal optimally uses a little of both (see Appendix 1). In this case monitoring and pay are complementary instruments and the new incentive contract is represented by

[4] Of course, this does not imply that the level of monitoring is efficient. As Bowles (1985) has pointed out, in the shirking model a slight reduction in monitoring accompanied by a slight increase in the wage yields a second–order loss to the principal and a first–order gain to the agent.

the contract B.

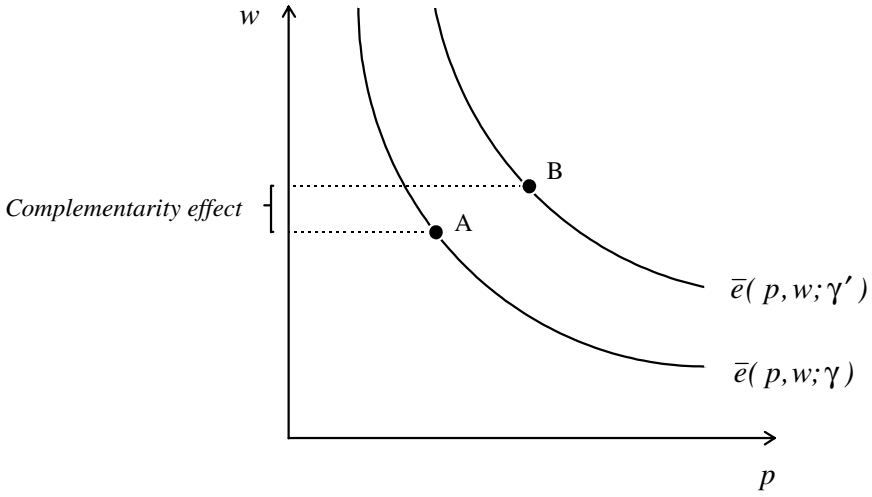

Figure 4

If the source of variation is β, nothing will happen with the equilibrium values of p and w provided that it is still profitable for the principal to remain in business.[5] Consequently, the discrete model yields a different implication for any possible source of variation.

Let us now investigate how the wage level and the accuracy of monitoring vary with the parameters β and γ in the case with continuous effort. The effect of a change in the principal's benefit from effort, β, is found by differentiating the two first–order conditions (8) and (9) to get the two equations

$$\begin{aligned} U^*_{ww}dw^* + U^*_{pw}dp^* &= -\frac{p^*}{\beta}r(e^*)d\beta, \\ U^*_{pw}dw^* + U^*_{pp}dp^* &= -\frac{w^* - w_0}{\beta}r(e^*)d\beta. \end{aligned}$$

Again, using Cramer's rule, we have that

$$\frac{dw^*}{d\beta} = \frac{[-p^*U^*_{pp} + (w^* - w_0)U^*_{pw}]r(e^*)}{[U^*_{ww}U^*_{pp} - (U^*_{pw})^2]\beta} \tag{20}$$

and

$$\frac{dp^*}{d\beta} = \frac{[-(w^* - w_0)U^*_{ww} + p^*U^*_{pw}]r(e^*)}{[U^*_{ww}U^*_{pp} - (U^*_{pw})^2]\beta}. \tag{21}$$

[5] See Appendix 1 for details.

The sign of $dp^*/d\beta$ is now easily determined. By the second–order condition and Assumption 1, we have

$$\operatorname{sign}\frac{dp^*}{d\beta} = \operatorname{sign}[-(w^* - w_0)U^*_{ww} + p^*U^*_{pw}] = \operatorname{sign} r(e^*) > 0.$$

The intuition is quite clear. When the principal has more to gain from increased effort, he will find it optimal to engage in more monitoring – this being one way of making sure that the agent does not shirk. A similar argument would seem to guarantee that the wage is increasing in β. However, this is *not* true without further assumptions. More precisely, we have from (20) that $dw^*/d\beta > 0$ if and only if

$$\mu p^* M''(p^*) + (w^* - w_0)r(e^*) > 0. \tag{22}$$

Using equation (9), an alternative statement of the condition is

$$-p^*\frac{M''(p^*)}{M'(p^*)} < 1.$$

Thus, a sufficient, but not necessary, condition is that $M(p)$ is convex. Generally, we may allow some increasing returns to monitoring. Only when the returns to monitoring increase sufficiently fast could increased profitability of effort lead to a switch away from monetary incentives towards monitoring. Note that if the principal can commit to random monitoring, the relevant monitoring cost function does become convex even if $M(p)$ is concave.

Finally, we investigate the impact of a change in the cost of effort. The formulas in this case become

$$\frac{dw^*}{d\gamma} = \frac{-p^*\mu M''(p^*) - (w^* - w_0)r(e^*)}{[U^*_{pp}U^*_{ww} - (U^*_{pw})^2]\gamma}U^*_{pw} \tag{23}$$

and

$$\frac{dp^*}{d\gamma} = \frac{-p^* r(e^*)}{[U^*_{pp}U^*_{ww} - (U^*_{pw})^2]\gamma}U^*_{pw}. \tag{24}$$

We see directly that, even here, monitoring and pay are complementary instruments if and only if condition (22) is fulfilled. The above results can be stated as follows.

Proposition 2 *If the source of variation is β or γ, monitoring and pay are complementary instruments if and only if $-p^*M''(p^*)/M'(p^*) < 1$.*

Since this is a very weak condition, we can safely conclude that w and p are complementary with respect to changes in β and γ. Indeed, as one would expect, the condition in Proposition 2 is weaker than the condition in Lemma 1; complementarity holds even more generally for variation in β and γ than for variation in μ. To see this, consider the condition

$$U_{ww}U_{pp} - (U_{pw})^2 > 0,$$

which must hold at (p^*, w^*) (this is one of the second–order conditions). A few computations reveal that the inequality can be written as

$$\mu p M''(p) + (w - w_0)r(e) > r(e)\left(\frac{-\gamma^2 r(e)}{\beta p h(e)} - (w - w_0)\right).$$

It is straightforward to check that the right hand side of this inequality is positive if and only if $U_{pw} > 0$. Hence, if $U_{pw} > 0$, (22) is implied by the second–order condition. This also completes the proof that Proposition 1 holds regardless of the source of variation.

5 Empirical implications

Prendergast (1999) argues that it is difficult to test for the existence of efficiency wages because we do not know if we would expect to see wages and supervisors as substitutes or complements. He writes (page 45): "The problem is that either may easily arise in a world of efficiency wages and depends critically on the source of variation across firms. On the one hand, if the source of variation across firms is the cost of supervisors, then the two instruments are likely to be substitutes, where firms substitute away from high-cost supervisors into wages. On the other hand, if the source of variation across firms is in the return to effort (so some firms value effort exertion more than others), those firms that want more effort will use more of both instruments relative to those that do not value such high effort."

Our analysis shows that one of Prendergast's conjectures is mistaken. Even if the source of variation across firms is the cost of supervisors, the two instruments are likely to be complements. Hence, the theory predicts that monitoring and pay move in the same direction regardless which parameter varies, offering some hope for empirical analysis. However, there are several important qualifications.

In theory, the only additional assumption we need in order to take the prediction to data is that the parameters $(\beta, -\mu, -\gamma)$ are affiliated random variables – see Holmström and Milgrom (1994). I.e., these variables should be

either independent or positively correlated.[6] While the affiliation assumption may look weak at first sight (we make the assumption of statistical independence quite routinely) it is unfortunately rather strong in the present context. To see this, consider a set of potential firms, whose attributes are drawn independently. Suppose that only firms which can cover a positive entry cost will enter. Then, to the extent that firms with high costs enter, these firms must also have relatively large benefits. In other words, firms with high μ or γ are more likely to have high β. Similarly, firms with high μ are likely to have low γ and vice versa: The parameters $(\beta, -\mu, -\gamma)$ are not affiliated for the firms which choose to enter, which means that there will be a sample selection problem, at least in markets with high entry costs.

Unfortunately, failure to address the sample selection problem is not the only weakness of existing empirical work.

5.1 The measurement of monitoring

The major drawback of most empirical work to date is that it uses monitoring intensity as a measure of monitoring accuracy, p. Monitoring intensity is proxied by the number of supervisors per employee or by the frequency of supervision. (Among the studies suffering from this weakness are Leonard (1987), Gordon (1990,1994), Kruse (1992), Neal (1993) and Arai (1994a, b).) But a large p does not necessarily correspond to a large number of supervisors. While a high p could be an effect of many supervisors, it could also be an effect of easy monitoring, in which case there may be very few supervisors. Conversely, there could be few supervisors either because monitoring is very easy or because it is very difficult.[7]

To be specific, suppose that each supervisor hour costs one dollar, and that the required number of supervisor hours needed to attain accuracy p is $\mu M(p)$. In other words, μ is a measure of how difficult it is to monitor. How does the optimal number of supervisor hours relate to μ? Differentiating and evaluating at the point p^* we have

$$\frac{d\mu M(p^*)}{d\mu} = M(p^*) + \mu M'(p^*)\frac{dp^*}{d\mu}. \tag{25}$$

We know from Section 2.3 that $dp^*/d\mu < 0$. Hence, the sign is ambiguous in general. However, we can say more. As μ tends to zero, the second term

[6] If there is positive correlation between the parameters, one obviously needs to be careful in making statements about the *magnitude* of complementarity between instruments. For an analysis of this issue, see Athey and Stern (1998).

[7] puts it, the number of supervisors captures the quantity but not all aspects of the quality of supervision.

vanishes (at least if $M'(p)$ is bounded above), so in this case the number of supervisor hours goes up as monitoring becomes more difficult. On the other hand, as μ tends to infinity, p^* tends to zero and the first term vanishes. Indeed, when monitoring is sufficiently costly, it is not worthwhile doing any monitoring at all. Hence, over some interval the number of supervisor hours must decrease as monitoring becomes more difficult.

We have thus proved that the number of supervisor hours is a non–monotonic function of μ. It follows that there will also be a non–monotonic relationship between the wage and the number of supervisor hours. Very low wages and very high wages should both be associated with few hours of supervision. The highest wages should be paid when monitoring is so accurate that very few supervisor hours are needed; the lowest wages should be paid when supervision is so costly that the principal cuts back on supervision for this reason. The big difference between the two cases is that in the former jobs, workers are expected to work very hard, and in the latter jobs they are not expected to work much at all.

If we think of primary sector jobs as jobs with high wages, our theory says that monitoring is easier in the primary sector than in the secondary sector. On the contrary, the "technical" assumption that there are only two effort levels led Bulow and Summers (1986) to make exactly the opposite prediction. However, the leading example of a primary sector job offered by Bulow and Summers is Henry Ford's five dollar day, the pay policy introduced by the Ford Motor Company in 1914, which represented almost twice the wage that similar workers could be expected to earn elsewhere. Reportedly, productivity increased by about fifty per cent, a feature which squares well with the model offered here. The key question is: did the change in pay policy relate to changes in supervision technology, and if so how? The common view is that mechanization of production and standardization of tasks generally made it easier to deter shirking. According to our theory, then, the increase in pay could be due to easier monitoring.

Given the above observations, we should not be surprised that earlier studies have been unable to find a strong relationship between hours of monitoring and the level of pay. In the terminology of our model, most of the empirical work consists of regressions of the wage level w or the wage premium $w - w_0$ on the monitoring cost $\mu M(p)$. Our analysis suggests that there will be (at least) two values of w for every value of $\mu M(p)$. Hence, it is invalid to estimate any functional relationship with w as the dependent variable. On the other hand, the model does not in general rule out the possibility of estimating $\mu M(p)$ as a function of $w - w_0$. A natural direction for future research is to reexamine existing data with the monitoring cost rather than the wage level as a dependent variable, and allowing for the kind

of non–linear relationship predicted by the theory.

5.2 Minimum monitoring restrictions

In some professions, employers are not free to decide the ratio of supervisors to workers. For example, Krueger and Groshen (1990) argue that regulation of working conditions makes the supervisor/staff ratio largely exogenous in the case of hospital nurses in the United States.[8] These authors go on to show that there is a negative empirical relationship between supervision and pay for hospital nurses. If effort is freely flexible, this finding contradicts our model.

Proposition 3 *Suppose the level of monitoring, p, is subject to binding regulations. Then the optimal wage, w^*, is an increasing function of p if and only if $U_{pw}(p, w^*) > 0$.*

PROOF: Suppose that regulations force p to be greater than the principal would have wanted. Hence, the principal has only one instrument: the wage. The solution to the principal's problem is given by the first–order condition $U_w(w^*, p) = 0$. Differentiating this condition gives

$$\frac{dw}{dp} = -\frac{U_{pw}(p, w^*)}{U_{ww}(p, w^*)}.$$

Since the denominator is negative by the problem's second–order condition, it follows that the wage should increase with monitoring if and only if $U_{pw} > 0$. ■

As we have argued above, the cross–derivative is indeed positive for reasonable specifications. If monitoring is in fact largely exogenous as the authors claim, the data used by Krueger and Groshen is immune to the criticisms raised above, because there will be a monotonic relationship between whatever is regulated and actual monitoring, p. Hence, their empirical finding seems to reject the present model.[9]

[8] Note that Krueger and Groshen are primarily concerned with an endogeneity problem that we have ruled out here: the possible substitution between supervisors and workers in the production function.

[9] It is worth noting that Krueger and Groshen appear to be aware that scale effects could possibly overturn the prediction that monitoring and pay are substitutes. They write (page 138-S): "Thus, holding workers' effort level constant, the efficiency wage model predicts that increases in monitoring would be associated with lower wages." But they then proceed to test the hypothesis without any controls for worker effort.

One objection to Krueger and Groshen's study is that they view wages at a point in time rather than the expected net present value of wages. If regulation forces a hospital to have more nurse supervisors, each nurse's probability of advancing to the rank of supervisor increases. Hence, average lifetime pay may be higher even if pay at the lowest rank is lower. Thus, we are not convinced that there is really a negative relationship between monitoring and pay in this case.

5.3 Direct measures of monitoring

Another paper which credibly identifies the quality of monitoring and not just the monitoring cost is Krueger's (1991) study of the fast–food industry. Krueger compares the level of pay in franchise outlets with the level of pay in company–owned outlets, finding that pay is lower in the former. He argues that there should be more monitoring in the franchise outlets, because there the residual claimant can monitor workers directly, and that this might explain the lower pay for non–supervisory staff in the franchise outlets. As shown in equation (13),[10] this conclusion is incorrect when the scale effect is taken into account.[11] Hence Krueger's finding contradicts the shirking model, rather than confirming it.[12]

[10] Krueger also argues that it is the different contractual arrangements that explain why franchisees expend more effort monitoring their workers than do managers of company-owned restaurants. He writes (page 76): "An owner-manager of a franchise has a strong incentive to expend effort supervising and monitoring his workers because he receives the residual profit generated by the enterprise; whereas a manager of a company-owned establishment is usually not paid a share of the establishment's profit, and his actions are not perfectly observed by his principal, the parent company." In our model this may be more appropriately captured by different βs rather than different μs. Hence, for those believing in the argument above, this reference should be to equation (20) instead.

[11] The fact that Krueger only looks at two levels of monitoring does not affect the qualitative conclusion. It is straightforward to show that our results carry over to the case in which the level of monitoring is chosen from a finite set P of alternatives: When the principal chooses to jump to a higher level of monitoring, at which point the discreteness of P implies that the new p is "too high", he will also choose to raise the wage. This follows directly from Proposition 2.

[12] A possible conventional explanation for Krueger's findings is that it is more pleasant to work at franchised restaurants. We see two potential reasons why this might be so. First, workers report that they find supervision to be more adequate at franchised restaurants; perhaps this also means that the workers are on better terms with their supervisors? Secondly, franchised restaurants tend to be somewhat larger and to be less centrally located. Both factors may contribute to a more even work load throughout the day.

6 Final remarks

We have shown that a very natural generalization of the conventional shirking model completely overturns previous intuitions. Moreover, most of the empirical literature fails to properly address the model's implications. In our view, only two studies of the relationship between monitoring and pay are appropriately designed. Rebitzer (1995) isolates the substitution effect, finding that monitoring and pay are substitute instruments for implementing a given effort level. However, as Prendergast (1999, p.44) notes, this prediction is not unique to efficiency wage theory. Krueger (1991), on the other hand, finds that monitoring and pay are substitutes even when both substitution and scale effects are present. As we have shown, this finding should be considered a rejection of rather than a confirmation of efficiency wage theory.

The apparent empirical failure of efficiency wage theory to explain the monitoring/pay relationship in Krueger's study begs the question whether other theories perform any better. Somewhat surprisingly, very few alternative theories exist. Holmström and Milgrom (1994) show that monitoring and pay/performance sensitivity are complements in their linear incentive model, a point which is also found in Milgrom and Roberts (1992, p226-7).[13] However, they do not study the covariance between monitoring and average pay, which is the central issue in the empirical literature that we address here. An alternative theory of the observed negative relationship between monitoring and average pay, which is put forward informally by Groshen and Krueger (1990), is that the return to monitoring is higher when workers have low skills. If this theory is correct, the negative relationship should disappear with better controls for worker skills.

[13] This model differs from the efficiency wage model in a number of respects. In the linear incentive model, agents are assumed to be risk averse, but to have unlimited liability. Thus, the mechanism behind the complementarity between monitoring and incentive intensity is quite different; more monitoring reduces the income risk, and thereby reduces the cost of strong incentives.

References

Arai, Mahmood (1994a): An Empirical Analysis of Efficiency Wages and Wage Dispersion, *Scandinavian Journal of Economics* 96, 31–50.

Arai, Mahmood (1994b): Compensating Wage Differentials versus Efficiency Wages: An Empirical Study of Job Autonomy and Wages, *Industrial Relations* 33, 249–262.

Athey, Susan and Scott Stern (1998): An Empirical Framework for Testing Theories About Complementarity in Organizational Design, NBER Working Paper no 6600.

Bowles, Samuel (1985): The Production Process in a Competitive Economy: Walrasian, Neo–Hobbesian, and Marxian Models, *American Economic Review* 75, 16–35.

Braverman, Harry (1974): *Labor and Monopoly Capital,* New York: Monthly Review Press.

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

Demougin, Dominique and Claude Fluet (1997): Monitoring versus Incentives: Substitutes or Complements? Mimeo. Université du Québec à Montréal.

Gordon, David (1990): Who Bosses Whom? The Intensity of Supervision and the Discipline of Labor, *American Economic Review, Papers and Proceedings* 80, 28–32.

Gordon, David (1994): Bosses of Different Stripes: A Cross–National Perspective on Monitoring and Supervision, *American Economic Review, Papers and Proceedings* 84, 375–379.

Groshen, Erica L. and Alan B. Krueger (1990): The Structure of Supervision and Pay in Hospitals, *Industrial and Labor Relations Review* 43, S134-S146.

Holmström, Bengt and Paul Milgrom (1994): The Firm as an Incentive System, *American Economic Review* 84, 972–991.

Krueger, Alan (1991): Ownership, Agency, and Wages, *Quarterly Journal of Economics* 106, 75–101.

Kruse, Douglas (1992): Supervision, Working–Conditions, and the Employer Size–Wage Effect, *Industrial Relations* 31, 229–249.

Leonard, Jonathan S. (1987): Carrots and Sticks: Pay, Supervision, and Turnover, *Journal of Labor Economics* 8, S135–S152.

Milgrom, Paul and John Roberts (1992): *Economics, Organization and Management,* (Englewood Cliffs, NJ: Prentice–Hall).

Neal, Derek (1993): Supervision and Wages across Industries, *Review of Economics and Statistics* 75, 409–417.

Prendergast, Canice (1999): The Provision of Incentives in Firms, *Journal of Economic Literature* Vol. XXXVII, 7–63.

Rebitzer, James B. (1993): Radical Political Economy and the Economics of Labor Markets, *Journal of Economic Literature* 31, 1394–1434.

Rebitzer, James B. (1995): Is There a Trade–Off between Supervision and Wages? An Empirical Test of Efficiency Wage Theory, *Journal of Economic Behavior and Organization* 28, 107–129.

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

Appendix 1

Suppose the effort level is restricted to take only the values $e = 0$ and $e = \bar{e}$. Provided that the principal wants to implement $\bar{e}$, his maximization problem is

$$\underset{p,w}{Max} \;\; \beta B(\bar{e}) - w - \mu M(p)$$

subject to the incentive constraint

$$p(w - w_0) = \gamma C(\bar{e}).$$

Using the constraint to substitute for w reduces the maximization problem to

$$\underset{p}{Max} \;\; \beta B(\bar{e}) - w_0 - \frac{\gamma C(\bar{e})}{p} - \mu M(p).$$

Assuming an interior solution, the first order condition is

$$\frac{\gamma C(\bar{e})}{p^2} - \mu M'(p) = 0, \tag{26}$$

which can be rewritten in terms of w as

$$\frac{w - w_0}{\gamma C(\bar{e})} - \mu M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right) = 0. \tag{27}$$

Total differentiation of (26) and (27) gives us the formulas

$$C(\bar{e})d\gamma - p^2 M'(p)d\mu - \mu p\left[2M'(p) + pM''(p)\right]dp = 0 \tag{28}$$

and

$$\mu C(\bar{e})\left[M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right) + \frac{\gamma C(\bar{e})}{w - w_0}M''\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)\right]d\gamma + \tag{29}$$

$$\gamma C(\bar{e})M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)d\mu - \left[2(w - w_0) + \mu\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)^2 M''\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)\right]dw = 0.$$

Using (28), it is easily seen that the effect of a change in μ on the monitoring accuracy, p, is

$$\frac{dp}{d\mu}|_{e=\bar{e}} = -\frac{pM'(p)}{\mu\left[2M'(p) + pM''(p)\right]}$$

and using (29) we get the analogous effect on the wage, w

$$\frac{dw}{d\mu}|_{e=\bar{e}} = \frac{(w - w_0)M'(p)}{\mu\left[2M'(p) + pM''(p)\right]} = \frac{w - w_0}{p}\frac{dp}{d\mu}.$$

Thus, this effect is always a substitution effect. And under the reasonable assumption that the denominator is positive in these expressions, we have the result that a decrease in μ will make the principal substitute wage for monitoring.

From the same formulas, (28) and (29), we find that the effect of a change in the parameter γ on p and w is

$$\frac{dp}{d\gamma}|_{e=\bar{e}} = \frac{C(\bar{e})}{\mu p\left[2M'(p) + pM''(p)\right]}$$

and

$$\frac{dw}{d\gamma}|_{e=\bar{e}} = \frac{C(\bar{e})\left[M'(p) + pM''(p)\right]}{p\left[2M'(p) + pM''(p)\right]} = \mu\left[M'(p) + pM''(p)\right]\frac{dp}{d\gamma}.$$

Thus here, given that $-pM''(p)/M'(p) < 1$ (inequality (22)), the effect is a complementary effect. Notice finally that there is no effect of a change in the parameter β on the equilibrium values of p and w, since β does not appear in either (26) or (27).