


Joseph A. Ritter is a research officer at the Federal Reserve Bank of St. Louis. Lowell J. Taylor is an associate professor at the Heinz School of Public Policy and Management, Carnegie Mellon University. Eran Segev and Joshua D. Feldman provided research assistance.



## Economic Models of Employee Motivation

Joseph A. Ritter  
Lowell J. Taylor

To most people it is a common sense proposition that hiring workers is a trickier problem than buying ballpoint pens. It is often difficult to find the right worker to hire, and workers who have already been hired can quit, steal, be hung over, refuse to cooperate with other workers, or simply not work very hard. In *some* workplaces *some* of these problems are relatively easy to solve, either by direct supervision or by directly linking pay to production. In general, however, things like ability, effort, and honesty are difficult to verify and consequently present special problems for personnel managers and economic theorists. The ways firms solve the problems of selecting, motivating, and retaining employees are potentially interesting to a wide cross-section of economists because they can affect how labor markets function and, therefore, how the entire economy operates.

This article presents an overview of economists' main hypotheses about the compensation strategies businesses use to address these kinds of problems. Broadly speaking, these solutions fall into three categories (with considerable diversity within each): piece rates, performance bonding, and efficiency wages. Piece rates link pay directly to workers' output. Performance bonding uses a combination of up-front payments from workers and conditional repayments to guarantee workers' performance. Firms that pay wages (we

use the terms "wage" and "compensation" interchangeably throughout the article) high enough to deter undesirable behavior by making a job too good to lose are said to pay efficiency wages.

It is fairly easy to see whether a firm is using some sort of piece rate plan. There is quite a bit of controversy, however, about whether firms that do not use piece rates adopt efficiency-wage or performance-bonding plans. We follow our overview with a discussion of the nature of the evidence supporting the different models.

### SIMPLE SUPPLY-AND-DEMAND MODELS OF LABOR MARKETS

On one level, economists can analyze labor markets using the same supply-and-demand model they might apply to, say, wheat. Supply increases as the price (wage) received by the supplier increases. Demand increases as the price paid decreases. Equilibrium occurs where supply equals demand.

For many purposes it is important to recognize that workers are not perfectly interchangeable; most nurses are not economists. This complication is easily handled by treating the markets for nurses and economists separately, each with its own supply and demand curves. Similarly, workers within the same profession are not typically interchangeable. An important dimension along which different kinds of workers can be distinguished is the collection of applicable knowledge and skills that economists call *human capital*. Levels of human capital vary not only across individuals, but also over time for a given individual. As an employee accumulates human capital, or as existing human capital deteriorates, the employee's compensation can be expected to change.

A worker's willingness to accept a particular job will be affected by agreeable and disagreeable facets of the job. Workers require a higher wage to accept a hazardous

job than a safe one. They may accept lower wages to work in a nice place, have flexible hours, or perform work that requires little effort. Differences in wages that come from these kinds of reasons are called *compensating differentials*.

The theory of labor demand is especially important for this article. The core of that theory is based on the observation that hiring an additional employee (or employee hour) will increase the profits of the firm as long as the employee's compensation is less than the value of the additional output the firm can produce after hiring the employee. The latter quantity is called the *value of marginal product* (VMP) and is calculated by multiplying the additional employee's marginal product by the price of the firm's product. This relationship defines the firm's labor demand curve. Since the marginal product is likely to decrease as the firm hires more labor (holding other inputs fixed), the firm's labor demand curve is downward-sloping: A firm that must pay higher wages will demand less labor. If there are no impediments, a labor market will reach equilibrium where supply equals demand.

The theory of supply and demand does a good job of explaining the broad outlines of labor markets, but a closer look reveals some cracks. This article concentrates on the fact that (unlike wheat, for example) the same worker behaves differently in different economic circumstances; the same worker might, for example, work hard at \$30 per hour but loaf at \$7 per hour. The simple supply-and-demand framework cannot encompass this possibility, so different kinds of models are needed.

## SPECIAL PROBLEMS IN LABOR MARKETS

A central task of economic theory is to boil a problem down to its essentials so that it can be thoroughly understood and carefully analyzed. In principle, after the core of the problem is understood, economists turn their attention to the nuances that separate their models (artificial economies) from reality. In the area of

worker motivation, labor economists have focused largely on three core problems: sorting potential employees, achieving optimal performance on the job, and regulating turnover.

### *Sorting Job Applicants*

In the textbook supply-and-demand approach to labor markets, sorting applicants is assumed to be a simple problem. That theory presumes that an employer knows how productive an applicant will be if he or she takes the job. An accounting firm (or anybody else) knows that an accounting major is likely to be a more effective accountant than a high-school dropout. But that kind of insight is only the tip of the iceberg and would not help to land an applicant a job in the accounting firm's personnel office. The difference between good and bad employees often depends on qualities that are difficult to discern (willingness to work hard, for example). If the firm designs the right incentives, however, it can encourage desirable applicants, even though the firm cannot easily identify them when they apply.

### *Performance on the Job*

Workers' behavior on the job can disrupt the firm's attempts to make money in many ways. A surly worker might drive away a customer. An employee who shows up late might make it difficult for other workers to do their own jobs. A worker might be careless or simply not work very hard. Workers might steal from their employer. The list is virtually endless.

Beyond the obvious, several aspects of these situations are important. First, none of the examples just mentioned is necessarily tied to any observable characteristic of a job applicant. Businesses use an arsenal of screening devices to try to avoid problems, but their effectiveness is manifestly limited. To get optimal performance from employees, the firm cannot rely on applicant screening alone but must also design effective incentives for existing employees.

Second, certain on-the-job problems are particularly critical because employees work together in most firms. A worker who shows up 10 minutes late, for example, is not a problem if his job is to sit in front of a computer and write articles, but if he works on an assembly line, he may force several hundred workers to start work 10 minutes late.

Third, the size and complexity of most workplaces make it impossible to detect all negative behavior. This fact triggers a powerful principle: To deter any behavior that is unlikely to be detected, punishment must be disproportionate. For example, suppose Joe likes hanging around the water cooler enough that he would be willing to pay the firm a dollar to be allowed this liberty for an extra hour each day, but his manager notices excess water cooler attendance only one time in a hundred.<sup>1</sup> To deter this behavior, then, the firm must impose a penalty greater than a dollar if Joe is caught hanging around the water cooler. This is, in our view, the central reason the basic supply-and-demand model is unlikely to be completely satisfactory in labor markets. The most severe punishment a firm can impose is firing, but in the basic supply-and-demand model, firing imposes very little cost on the worker, because the model assumes that markets are anonymous and function quickly and efficiently. Thus a terminated worker has no difficulty in finding a comparable job.<sup>2</sup>

## Turnover

Turnover can be very costly to the firm for two reasons. First, isolating and hiring a new worker can cost thousands of dollars for some jobs. Firms that outsource *part* of this activity to “head-hunters” (presumably because they think it is cheaper than doing it themselves) typically pay a commission that is a substantial fraction of the new worker’s annual pay. Second, new workers almost always need to accumulate some knowledge specific to the new job (firm-specific human capital). This process may require explicit training, or it may just mean that the new worker will not be fully

productive for some time. To the extent that firms cannot shift these costs to the new worker (through probationary wages, for example), firm-specific human capital is costly to the firm. Quit rates are not entirely outside the firm’s control, however. Compensation policies can provide incentives for workers to stay on the job.

## The Agency Problem

All of the problems mentioned in this section are corollaries of the maxim, “If you want something done right, you have to do it yourself.” The problem confronting the owner of a business is how to design incentives that will induce the workers to do it right or, more precisely, to behave in a way that maximizes the firm’s profits. This is an example of what economists call an *agency problem*: A principal (in this case the firm’s owner or manager) designs incentives for an agent or agents (the workers), who take actions that affect the principal’s well-being. The agency problem stems from the fact that there is a different connection between the agent’s actions and well-being than between the agent’s actions and the principal’s well-being.<sup>3</sup> For example, it is in the firm’s interest for a worker to work hard (the action preferred by management), but the worker may prefer to spend the morning at the water cooler.

## PIECE RATES

If the agency problem is related to the worker’s productivity, there is an obvious approach to solving it: Establish a direct connection between the worker’s output and his compensation. Many workers are compensated in ways that resemble piece rates: garment workers who are paid on the basis of output, sales workers paid on commission, auto mechanics in large dealerships whose pay is partly on a per-repair basis, and agricultural workers whose pay depends on the amount of fruit picked or rows of grape vines pruned.

One pervasive problem in firms that tie pay closely to some objective measure of output is that they often get exactly

<sup>1</sup> Perhaps because the manager is usually golfing. See footnote 3 below.

<sup>2</sup> Recent work on the dynamics of labor markets uses matching models in which the firm and worker bargain over gains generated by a good match. If the worker’s share is small, firing costs the worker little. Even when the worker’s share is larger, so that termination is a significant penalty, its credibility as a disciplinary device is limited because firing is costly to the firm too.

<sup>3</sup> The owner(s) of a large firm face another agency problem: how to get the manager of a firm to act in the interest of the owner(s). This problem has also been extensively studied under the heading of executive compensation. See Jensen and Murphy (1990).

what they pay for: behavior that changes the measure of output rather than output itself (Baker, Gibbons, and Murphy, 1994). Fraud and accounting tricks often allow employees to manipulate the output *measurement* without changing output. Or, perhaps worse, easily observable quantity may rise at the expense of less apparent quality. The dilemma is summarized by Gibbons (1996): “When measured performance omits important dimensions of total contribution [to the firm], firms understand that they will ‘get what they pay for,’ and so may choose weak incentives in preference to strong but frequently dysfunctional incentives.” In other words, firms facing these types of distortions may choose to use incentive systems that are less direct and less precise than piece rates.

The biggest impediment to the implementation of piece rates is that the output of individual workers is not easily measured in many jobs; reasonable, objective measures of performance do not exist. One reason is that it is usually difficult, if not impossible, to separate a particular worker’s performance from the overall performance of a group or firm. Inadequate output measurement makes piece rates far less effective.

Firms’ motives for using “weak” incentives can be even deeper than obviously defective or easily manipulated measurement systems. Holmstrom and Milgrom (1994) argue that when workers perform several tasks, incentives must be finely balanced to ensure that all the tasks get adequate attention. But if, for example, one task is easy to measure and another, equally important, task is hard to measure (cooperation, for example), it will be impossible to use “strong” incentives—piece rates—to motivate performance on the first task without also inducing the worker to neglect the second task. Sometimes, therefore, firms may forego the opportunity to use piece rates (or use only weak ones), even when they would ostensibly be feasible and effective. Holmstrom and Milgrom conclude that “the use of low powered incentives within the firm, while sometimes lamented as one of the major disadvantages

of internal organization, is also an important vehicle for inspiring cooperation and coordination.”

Although a supervisor may be able to judge whether the worker is doing a good job over some period of time (we choose fuzzy words deliberately) and set pay accordingly, for two reasons this approach is not really a piece rate. First, evaluation by supervisors breaks the tight relationship between performance and pay that true piece rates can achieve in a simple environment.<sup>4</sup> Second, it introduces a time dimension to the relationship between work and compensation that changes it in fundamental ways from the simple immediate reward system of piece rates. The remaining approaches discussed here stress this time dimension.

## PERFORMANCE BONDING

In the face of workers’ inclinations to do various things contrary to the best interests of the firm, it is useful to divide compensation in two pieces. One piece is the level of compensation that the worker requires before agreeing to work for the firm at all. This piece includes any compensating differentials the firm must pay. For the next three paragraphs (only) we will call this component the base wage. The second piece of compensation convinces the worker to perform optimally—to work hard, stay sober, be unlikely to quit, and so forth. For the moment we will call this the bonus. If piece rates were feasible, this bonus could be zero. It might also be zero if it is easy to monitor the worker’s performance in relevant ways. As we argued above, these cases are far from universal.

The base wage does not help motivate the worker, because it simply measures the alternative value of his time. It does not motivate him to do things he is disinclined to do (work hard, for example). Compensating differentials reflect the market’s valuation of things such as high effort, but if the employer cannot perfectly monitor the employee’s behavior, a compensating differential will not ensure that high effort

<sup>4</sup> The literature on incentive pay in economics studies the limited extent to which efficient employment relationships can be achieved in the face of this kind of slippage. A nice summary of this and other issues in incentive pay is Gibbons (1996).

is forthcoming. Clearly, the bonus, also, will not motivate workers if it is not conditional on performance in some way. So the firm must have a scheme whereby the worker is periodically evaluated and receives the bonus only if the evaluation suggests that his performance exceeds some threshold. Suppose that the evaluation is reasonably honest and accurate (closely related to the worker's actual performance). If the bonus is big enough, it will provide adequate incentive for the worker to perform as the firm wants. How big it needs to be will depend on how likely it is that the firm's evaluation will detect suboptimal performance.

There is a flaw in this plan, however: The worker's compensation (base wage plus bonus) may exceed the value of his marginal product if the bonus is too large. In this case, firms could simply decide it is not profitable to hire workers whose compensation exceeds the value of their marginal product and make no further effort to solve the agency problem. But here is a better idea: The firm could require the worker to give it some money at the beginning of the evaluation period and promise to pay it back with interest at the end, conditional on adequate performance. Now the firm is free to hire workers up to the point at which the value of the marginal product of labor equals the base wage because the workers are paying their own bonus. In economics jargon, they are posting a bond to guarantee their own performance. The firm still must compensate workers to do things they do not want to do (pay a compensating differential, in other words), but the bond guarantees that the firm will get what it pays for (if the bond is large enough to offset whatever temptations cause the firm's agency problem).

At first this idea appears to be a case of economic theory run amok. Jobs that require an explicit bond, as just described, are extremely rare, and this seems to be conclusive evidence against this theoretical approach. Indeed, Carmichael (1989) is blunt about this fact: "I know of no labor markets anywhere in the world or in history where this practice has been

widespread." But to write the idea off would be to underestimate the ingenuity of economic theorists.

### Work-Life Incentives

Edward Lazear (1979, 1981, 1995) has argued that actual compensation plans implicitly use the bonding idea and, moreover, that recognizing this fact can help to explain some features of labor markets that otherwise appear quite odd. Lazear's basic insight is that if firms and workers have full access to capital markets—that is, if they are able to save and borrow effectively—neither side should care whether workers' compensation exactly equals the value of marginal product (VMP) on any given day. Instead, both care about the present value of wages and VMP over the working life of the employee. This observation suggests a new strategy that makes the performance bond an implicit part of compensation, rather than an explicit up-front payment. Lazear (1995) calls this approach "work-life incentives." The same idea goes by various names, including "life-cycle incentives" and "upward-sloping age-earnings profiles" or "tenure-earnings profiles."

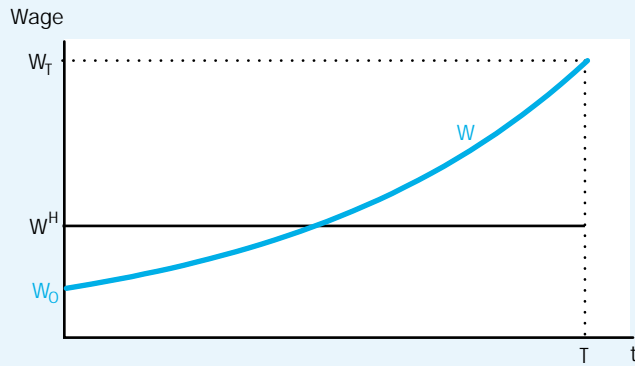
Lazear poses a simple agency problem: Suppose workers can work at either a high or a low effort level, and that they are indifferent among the options of working hard for wage  $W^H$ , working at a low effort level for wage  $W^L$ , or not having the job.<sup>5</sup>  $W^H$  and  $W^L$  are the workers' high- and low-effort *reservation wages*. (In reality, of course, firms must decide on an acceptable effort level, but adding that decision would not substantially change any part of this article.) Their difference,  $e$ , measures the monetary value to the worker of the extra effort. Employees who work hard are more productive than those who supply low effort, so  $VMP^H > VMP^L$ . Suppose that workers' productivity at each effort level does not change during their lifetimes. A firm that could be sure its workers were working hard would pay  $W^H$  and hire additional workers until  $VMP^H$  fell to  $W^H$ . A firm that knew its workers to be shirkers

<sup>5</sup> This effort-supply problem is frequently used because, despite its extreme simplicity, it captures ideas like Lazear's that are valid for many different and more complex situations.



Figure 1

## Work-Life Incentives Wage Profile



would pay  $W^L$  and hire until  $VMP^L = W^L$ . Some firms will choose the latter strategy, but if high effort is worth more to a particular firm than to workers ( $VMP^H - VMP^L > e$ ), the firm will want to choose a compensation mechanism that persuades the worker to work at the high effort level. These are the firms with agency problems.

For the reasons discussed above, paying workers  $W^H$  throughout their careers will not by itself convince them to work hard, even though their pay includes a compensating differential for high effort. Even a threat of termination would do no good, because their next best option is just as desirable as a high-effort job at wage  $W^H$ ; that is what we mean by a reservation wage. In other words, *the job itself has no value to the worker*. A firm following this strategy gets low output for high wages, a losing proposition.

Lazear observes that there is a simple way to make the job valuable to the worker. Consider the lifetime wage profile labeled  $W$  in Figure 1, which has been tilted so that the present value of wages paid on  $W$  between hiring at date  $t=0$  and retirement after date  $t=T$  equals the present value of a constant wage  $W^H$ , that is,

$$\sum_{t=0}^T \left( \frac{1}{1+r} \right)^t W_t = \sum_{t=0}^T \left( \frac{1}{1+r} \right)^t W^H,$$

where  $r$  is the interest rate. What happens to the difference between the present value of  $W$  and that of  $W^H$  as time passes? The

difference between them at any time  $s$  between hiring and retirement is

$$\sum_{t=s}^T \left( \frac{1}{1+r} \right)^{t-s} (W_t - W^H),$$

and is shown in Figure 2. This quantity is the value of a job that pays wages  $W_t$  from  $t=s$  until  $t=T$ . In Figure 2, the difference first rises as the initial negative  $W_t - W^H$  terms get dropped off the beginning of the sum, and the positive ones get less discounting because they are not so far in the future (the term  $[1/(1+r)]^{t-s}$  gets bigger as  $t-s$  gets smaller). Eventually, however, the terms getting dropped off the start of the sum are positive, and there are fewer and fewer terms to sum, so the difference falls. By retirement, the difference falls to  $W_T - W^H$ .

At any point during his working life, a worker who chooses to work at low effort gets a utility gain  $e$ , but gambles that he will be caught (with probability  $d$  for detection) and lose a valuable job.<sup>6</sup> This will be a good bet; that is, a risk-neutral worker will shirk, if<sup>7</sup>

$$(1) \quad e > d \sum_{t=s}^T \left( \frac{1}{1+r} \right)^{t-s} (W_t - W^H).$$

In Figure 2, the worker will work hard up to time  $s^{**}$ .

By adjusting the slope of the  $W$  wage path (but leaving its present value unchanged), the firm can make  $s^{**}$  equal  $T$ , thus giving workers incentives for adequate performance most of the time.<sup>8</sup>

Deferring compensation, as work-life incentives do, also discourages quits among current employees. An employee does not receive full compensation for past work until the end of his career; as a result, the job continues to have value and there is always an incentive to hang on a little longer. For a similar reason, work-life incentives also help to screen out applicants who, for one reason or another, would be more likely to quit: A worker who takes a job for just a year or two at a firm that uses work-life incentives is underpaid, since wages are initially below  $W^H$ .

<sup>6</sup> Of course, the firm has some control over  $d$ . It should be understood here as a stand-in for how difficult in general it is to monitor an employee's performance.

<sup>7</sup> A risk-neutral worker is indifferent between accepting and rejecting a fair bet. A risk-averse worker would require a bigger gain from shirking to accept a given risk to his job. Because the worker does not lose  $W_s$  if he shirks and is caught in  $s$  (he gets paid up until the day he is fired), the wage profile must still be sloped a little bit even when  $d=1$ .

<sup>8</sup> The worker always has an incentive to shirk in  $T$  because there is no stream of future payments left to lose. The firm could use a pension paid after  $T$  to give the job value in  $T$  (and before) as long as it could take the pension away up to the very last minute, if necessary.

It is easy to see that work–life incentives can solve a broad range of agency problems. In fact, in principle this approach can be applied simultaneously to every agency problem mentioned in the introduction except encouraging better job applicants. The reason Lazear’s approach is so versatile is that it works entirely by making the worker’s job valuable enough that he will not risk dismissal; it does not matter how you interpret  $e$ , as long as the expected loss from undesirable behavior [the right-hand side of (1)] is larger.<sup>9</sup> Tournaments and efficiency wages function in the same way. Why can’t Lazear’s approach help improve a firm’s applicant pool? The job does not have value until the worker posts bond, which happens *after* hiring.

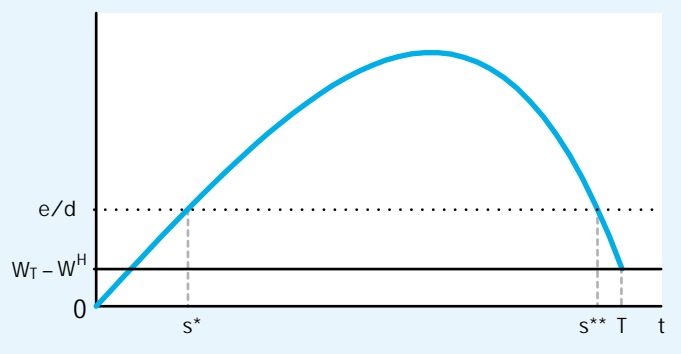
As a careful look at Figure 2 reveals, the agency problem is not completely solved even when the profile is adjusted so that  $s^{**} = T$ , because the value of the job is created by the accumulation of deferred pay, starting at zero. Initially, therefore, the value of the job is less than  $e/d$ , so some mechanism other than work–life incentives must be used to motivate workers during this interval.<sup>10</sup> The firm could require the worker to post an explicit bond at the beginning. But that would remove a major attraction of Lazear’s theory, that it does not require outright (net) payments from workers to firms, which are rare.

There is one last problem to wrap up. Since wages are at their highest late in life in Lazear’s model, workers have an incentive to hang on past  $T$ . The firm does not want this to happen because these high wages do not correspond to high current productivity; they are deferred compensation for past productivity. But this is not a flaw, it is a feature. Lazear (1979, 1995) observed that this “problem” could serve as an explanation of widespread mandatory retirement policies—policies that force employees to retire at a certain age, regardless of their productivity.

Mandatory retirement policies are now illegal for most workers in the United States, but Lazear (1995) shows how defined-benefit pensions (pensions that promise a set

Figure 2

## Difference Between Present Values of Wage Profiles



monthly benefit, based on years of service and rate of pay) can also be structured to bring about timely retirement. Decreasing life expectancy (as the worker ages) causes the present value of any given benefit level to decline as retirement age increases. Since the firm sets the rate at which benefits increase with years of service, it can therefore determine the age at which the present value is maximized. If the worker chooses to work past the age preferred by the firm, the present value of his pension starts to decrease, even if the monthly benefit level is still increasing. The worker is thus given a strong financial incentive to retire at the age preferred by the firm.

Another empirical implication of Lazear’s model, perhaps obvious enough to escape notice, is that earnings profiles slope upward throughout a worker’s career, even for workers who do not change jobs. This matches what labor economists have found in data on individual earnings histories. The upward-sloping earnings histories in the data do not seem to be fully explained by increasing productivity (human capital) as workers accumulate experience (Medoff and Abraham, 1980). Lazear’s analysis provides a supplementary reason for earnings to increase with experience.<sup>11</sup>

## Tournaments

Malcomson (1984) developed the idea that the internal hierarchy of a firm can be used as an effective incentive system.<sup>12</sup> In

<sup>9</sup> There are probably limits on how large the right-hand side of (1) can be made in practice. We discuss these at the end of this section.

<sup>10</sup> Akerlof and Katz (1989) pursue the implications of this observation. The problem disappears if we assume, as Lazear (1979, 1981) does, that shirking is detected with certainty at each instant in a continuous-time model. In that case the parameter that corresponds to  $d$  would be infinite.

<sup>11</sup> Details of the nexus between seniority and wages are surveyed by Hutchens (1989).

<sup>12</sup> Lazear and Rosen (1981) initiated the study of tournaments in labor economics. Studying the incentive effects of tournaments of the form, “The employee of the year will get a trip to Hawaii,” they showed that in some circumstances such tournaments could replicate the outcome of a piece-rate system, but with the advantage of needing information only about the relative rankings of workers, instead of their absolute productivity levels.

this type of model, a worker enters the firm at some level in a pyramid of possible jobs. The jobs at higher levels in the pyramid are rarer and pay more than those at the entry level. Periodically, the firm will promote a fraction of the employees from each level according to their ranking in some evaluation process, so that jobs at higher levels in effect become prizes in an ongoing tournament. The firm may supplement the prizes with terminations for employees who are not promoted. The chance of moving up and the competition needed to do so provide strong incentives for good performance.<sup>13</sup>

In a way similar to the work–life incentives described in the previous section, the size and number of prizes and penalties are set up so that the expected present value of compensation during a worker’s career equals the present value of his reservation wage. The incentives then operate in almost exactly the same way as work–life incentives: When the worker enters the hierarchy, he is initially paid less than the value of his marginal product (thus accumulating a bond). Wage increases are not certain in this model because luck (the quality of co-workers, for example), as well as effort, can influence success, but his expected lifetime earnings profile slopes up. High expected future income comes from a chance at promotions rather than increasing pay in the current job (as in Lazear’s model). Our comments in the previous section about quits apply here too.

Tournaments have some problems similar to the “you get what you pay for” problem that plagues piece rates. Because promotions are based on relative evaluations, workers may collude to reduce output (though, as in other cartels, this strategy is prone to defections) or spend time sabotaging each other’s chances for promotion rather than working.

From an economist’s point of view, the idea of hierarchies as incentive systems shares an attractive feature with work–life incentives. The logic of work–life incentives simultaneously solves an agency problem and provides an explanation for mandatory retirement, a phenomenon that had puzzled

economists. Similarly, tournament models provide a workable solution to an agency problem, and they help explain why hierarchies exist at all, why firms often prefer to promote existing employees rather than to hire new ones, and why the variance of earnings within an organization is greater for employees with more seniority.

### *Problems with Performance Bonding*

Few economists would dispute that mechanisms like those described in this section exist, and that managers of firms are aware of and try to exploit the incentives that the mechanisms provide. Controversy arises over whether compensation schemes based on the bonding principle can be pushed far enough to solve completely the motivation problems that firms face.

This controversy is important because bonding models allow firms to solve their agency problems at no cost and without altering the basic principles of supply and demand in the labor market. Although these models break the tight link between wages and value of marginal product, firms still end up equating the two, but they are averaged over a worker’s lifetime or across workers who enter a tournament. Therefore, *ex ante* decisions are not affected by the use of performance bonding. Bonding produces, in economists’ jargon, *first-best solutions*. If first-best solutions exist—that is, if bonding approaches can fully solve the agency problem—they will presumably be firms’ preferred approach. Barriers to their use open the door to second-best solutions like efficiency wages, which are considered in the next section.

The most important criticisms of performance bonding fall into four categories: imperfect financial markets, legal barriers, cheating (moral hazard) problems, and problems that come from hidden information. Explicit bond posting is rare in labor markets. In fact, it is unusual to see firms taking *anything* other than the job itself from workers who are fired. In other words, to the extent that bonding arrangements are used by firms, the value of the bond is somehow embedded in the job itself.

<sup>13</sup> The problem of motivating the individual(s) at the top of the hierarchy remains. This is, again, the problem of executive compensation mentioned in footnote 3.



Understanding why explicit performance bonds are hardly ever used obviously helps explain why firms might choose roundabout practices like work-life incentives and tournaments. The near impossibility of using explicit performance bonds is also important because there are limits to the implicit bonding schemes Lazear and others have proposed. For example, it is easy to construct examples in which adequate work-life incentives (steep enough wage profiles) require negative wages early in a worker's career. In other words, it is not always possible to tilt the wage path enough to get high effort *and* avoid explicit payments from workers to firms. Some other mechanism, such as efficiency wages, may therefore be necessary to change workers' behavior sufficiently. So why are explicit performance bonds so rare?

One reason may be that workers who are just starting a job have difficulty coming up with enough money to post an explicit bond. This conjecture challenges the assumption that workers can lend and borrow freely. Instead, they are liquidity constrained; they can save (lend), but their borrowing ability is limited.

Dickens et al. (1989) discuss a second reason explicit bonds may not be a useful option: There are limits on the types of contracts that governments will enforce. In particular, under American and English common law, courts refuse to enforce contract provisions they interpret as penalties (as distinct from damages). When the probability of detecting workers' misbehavior is low, performance bonds must be large because the disincentive to workers comes from the expected loss (the size of the bond times the probability of losing it), not the actual loss. Courts will typically not enforce contracts in which workers forfeit bonds that are disproportionately large. The courts do not, however, view firing as a penalty in this sense. Therefore *implicit* bonding arrangements are not limited by this legal standard. Implicit bonding also does not require explicit enumeration of the types and quantities of undesirable behavior that will result in penalties. Explicit contracts

would be limited to a relatively small set of legally verifiable actions.<sup>14</sup> The remaining problems with performance bonding apply to implicit bonds as well.

In addition to the common-law legal principle just mentioned, many countries have laws that interfere with the use of performance bonds. In the United States, for example: (1) mandatory retirement is illegal for most workers; (2) minimum wage laws interfere with firms' ability to pay very low wages to workers at the start of their careers; and (3) employers are required to vest workers in defined-benefit pension plans after five years. (This makes the job less valuable because it separates claim to a pension from continuation of the job.)

One problem with performance bonds that stands out in most people's minds is cheating by the firm, an example of moral hazard. If the worker's performance were objectively verifiable, the employer could probably use piece rates or something like them. In most jobs, though, performance is judged, somewhat subjectively, by management. This gives the firm a clear incentive to misrepresent the worker's performance in order to keep the bond. This is a compelling argument, but there are some considerations that mitigate it.

First, since other workers usually have their own subjective evaluation of a worker who is fired, firms that regularly exploit this opportunity may develop a bad reputation. If either existing workers or new applicants recognize that there is a substantial chance that they will lose their bond even if they perform well, the bond no longer provides the desired incentive. In addition, workers would require compensation in some form, probably higher wages, for the expected loss of the bond.

Second, promotion tournaments avoid the problem to a certain extent in the following way: If firms use a fixed number of prizes that will definitely be awarded according to the relative rankings of existing workers, the firms have no incentive to cheat. If they must fill the slots anyway, they are happy to fill them with the best workers. Of course, firms have

<sup>14</sup> Hart (1995) contains a very persuasive discussion about the practical (and, thus, legal) limits of legal contracting.

an incentive to avoid filling the slots at all (that is, awarding prizes) unless they serve some further function in the organization, but this ploy is easily observed by workers, so it would quickly destroy the incentive effects of the tournament. Ritter and Taylor (1997) argue that seniority-based layoffs have a similar advantage. Lower layoff probabilities for more experienced workers result in an upward-sloping experience-expected earnings profile, like that achieved by tournaments, even if the profile of actual wages is flat. The firm does not care which workers it lays off, since each is paid a wage equal to the value of his marginal product. It thus has no incentive to cheat.

The final category of criticism is based on two principles: (1) that workers will insist on competitive rates of return on the bonds they post and (2) that the firm has better information about the rates of return that workers will actually receive than do the workers. If a business shuts down, workers who have posted bonds through low wages early in their careers lose the entire value of their bond. Similarly, if a firm hits a rough spot and responds by eliminating higher-level positions to make itself more competitive, prizes are removed from its promotion tournament, lowering the expected payoff to the bonds that workers posted by accepting low wages in entry-level positions. The first principle implies that workers will expect to be compensated for events like these. The second says that firms have private information about how likely these events are.<sup>15</sup> Ritter and Taylor (1994) show that in these circumstances, risky firms (where workers insist on a higher rate of return on their bonds) have an incentive to pretend to be safe firms so that they can pay lower rates of return on the bonds. Workers, unable to distinguish between the two, require a rate of return above what they would demand from known safe firms. This makes performance bonding costly and, therefore, undesirable for safe firms, which separate themselves from risky firms by paying efficiency wages.

## EFFICIENCY WAGES

Bonding mechanisms like work-life incentives and tournaments can provide an effective resolution to the agency problem because they make jobs valuable to workers. Workers have an investment for which the return is tied to continuation of the job. They are therefore less likely to quit or to take actions that would result in their dismissal.

How should firms proceed if any of the economic or institutional reasons discussed above limit their use of performance bonds? The most obvious solution is to make jobs valuable in a direct manner—by paying more. The firm's strategy here entails the use of a "carrot" and a "stick." As in the work-life incentives model, the stick is the threat of dismissal. The carrot is the promise of a high-paying job.

To see how this works, we return to the simple effort-supply problem that motivated our discussion of work-life incentives. We assumed for simplicity that workers could either work hard (high effort) or shirk (loaf). Workers have reservation wages  $W^H$  and  $W^L$  for high and low effort levels, which are related by  $W^H = W^L + e$ . We call  $e$  the difference in effort levels, but it is really the amount of money that makes the worker indifferent between high and low effort.

Each day, a worker must decide whether to work hard or loaf. The consequences of this decision mirror those in the work-life incentives model: If he loafs, he gets immediate gratification worth  $e$ , but the probability is  $d$  that he will get caught, be fired, and lose a series of wages that exceed his reservation wage. Thus he will loaf at time  $s$  if

$$(2) \quad e > d \sum_{t=s}^T \left( \frac{1}{1+r} \right)^{t-s} (W_t - W^H).$$

It is not a coincidence that (2) looks the same as (1); they express the same gamble for the worker. There are, however, two differences that are not immediately apparent from the inequality alone. In the work-life incentives model, the worker

<sup>15</sup> The information is private in the technical sense that only the firm knows it, and it cannot be costlessly verified if the firm chooses to reveal it. The latter condition gives firms an opportunity for strategic misrepresentation.

posts a bond by accepting  $W_t < W^H$  early in the his career, so (1) does not hold at that stage. A premise of the efficiency-wage literature is that, for one reason or another, bonds cannot fully solve the agency problem. The crudest efficiency-wage models assume they cannot be used at all. In (2), therefore,  $W_t \geq W^H$  all the time, but in (1),  $W_t < W^H$  at the beginning of the worker's career.

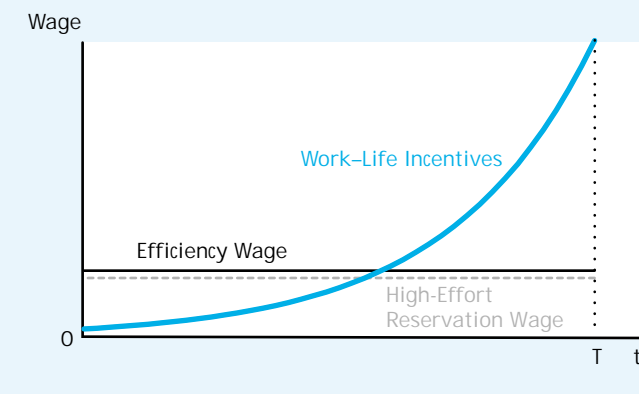
An important consequence of not allowing a bond to be posted is that it is impossible both to solve the agency problem and to match the present value of a worker's lifetime pay with the present value of his reservation wage; to solve the agency problem, the firm must pay an efficiency-wage premium. The efficiency wage is the lowest wage that will induce high effort, that is, the wage that would make (2) into an equality. Because the wage premium reduces profits, paying efficiency wages would be a second-best solution for the firm if some form of performance bonds could be used. Because it must pay a wage premium, an efficiency-wage firm demands less labor and produces less output than an otherwise identical firm that has no agency problem (or can solve its problem with performance bonds).<sup>16</sup> Our formulation in (1) and (2) highlights that the problem might be solved by some combination of performance bonds and efficiency wages, depending on how far the firm can push performance-bonding strategies.

The second subtle difference between (2) and (1) is in the interpretation of the reservation wage, and it arises because of the possibility of involuntary unemployment. We postpone discussing this until the next section.

The sum in (2) is the present value of efficiency-wage premiums—the value of the job relative to the reservation wage. To deter shirking, the firm must set the wage high enough to make the present value at least as great as  $e/d$ . Wages any higher than that would cut unnecessarily into profits. Thus the value of the job must always equal  $e/d$ .<sup>17</sup> Figure 3 shows the lifetime wage profiles that come out of the efficiency-wage and work–life incentives

Figure 3

## Work–Life Incentives and Efficiency-Wage Profiles



models using the same  $W^H$ ,  $e$ , and  $d$ .<sup>18</sup>

How do the solutions shown in Figure 3 change as the situation changes (across firms, for example)? First, if monitoring is more difficult ( $d$  is smaller) or more effort is required ( $e$  is higher), the efficiency wage will rise; larger carrots must be dangled to achieve optimal performance. In performance bonding models, bigger bonds are necessary—workers must give the firm larger carrots to be dangled in front of them. In the work–life incentives model, this means that the wage profile must be steeper, since the bond is accumulated during the phase in which the worker is underpaid. (For the same reason,  $s^*$  increases.) A fall in  $d$  works on the cost side of the worker's mental calculus. He recognizes that the chances of “getting caught” have fallen, and therefore a bigger penalty is required to induce him to forego a gain of  $e$ . An increase in  $e$  is simply an increase in the benefit of loafing and requires a more valuable job, so again the wage profile is steeper.

Suppose there is always a chance that the job will end for reasons unrelated to performance. The worker's wife could get an attractive job in a different city or the firm could shrink. We have not built this wrinkle into our simple versions of the models, but it is easy to apply the logic of the previous paragraph to see how this consideration affects the solutions. It all works through the value of the job. If a job separa-

<sup>16</sup> Typically, efficiency-wage models assume that the firm still operates on its neoclassical labor demand curve; that is, it hires labor until the VMP equals the wage. Its equilibrium VMP is thus higher than the equilibrium VMP of an otherwise identical firm with no agency problem.

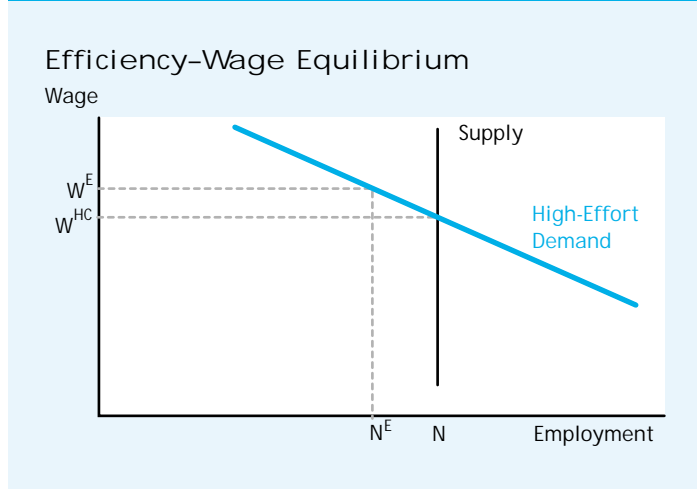
<sup>17</sup> This means that effort-regulation efficiency-wage models share the problem of inducing high effort in  $T$  (when there are no future wage premiums) mentioned in footnote 8. Similar strategies would solve the problem. The problem does not arise in many of the other types of efficiency wage models described below.

<sup>18</sup> The slope of the work–life profile in Figure 3 is set so that (1) holds with equality at  $T$  ( $s^{**} = T$ ). The efficiency wage path, like the work–life profile, assumes that the incentive problem is solved somehow in  $T$ . That being the case, the value of the job is always  $e/d$ , which gives an efficiency wage of

$$\left(\frac{r}{1+r}\right) \frac{e}{d}$$

in every period.

Figure 4



tion is more likely, there is a larger chance that the worker will never see some of the high wages promised in the future. This factor reduces the value of the job, so the firm must either pay a higher efficiency wage or require a larger bond. Using this reasoning, a firm that finds itself paying efficiency wages might also find it profitable to offer relatively stable employment, since stable employment would reduce the efficiency wage. Such a firm would sometimes operate off its VMP curve.

### Efficiency Wages and Unemployment

We have described efficiency wages from the standpoint of a single firm. When Shapiro and Stiglitz (1984) first introduced a close relative of the efficiency-wage model presented above, their primary focus was on the implications of this model of compensation for unemployment rates. This section presents the core of their argument.

Suppose there are lots of identical employers, each facing the same agency problem—encouraging high effort. There are also  $N$  workers who each supply one unit of labor, inelastically. If there were no agency problem, labor could be bought and sold like wheat. The applicable supply-and-demand graph would look like Figure 4. Suppose, as we have throughout this article, that the marginal product of effort is so high

that firms always prefer to pay for high effort. The competitive equilibrium wage is  $W^{HC}$ , where supply equals demand. This is also the high-effort reservation wage; any worker paid less than  $W^{HC}$  would immediately move into a comparable position with another firm. No worker would care about losing his job or whether his next job was working hard for  $W^{HC}$  or not so hard for some low-effort reservation wage (not shown).

Now introduce the agency problem. For the reasons given in previous sections, all workers would shirk at wage  $W^{HC}$ , and firms would not be getting their money's worth. In efficiency-wage models, firms make jobs more valuable to deter shirking. They do this individually by raising wages and reducing their own labor demand. Although they do not collude, their actions move them collectively up the high-effort demand curve to wage  $W^E$ , where they employ only  $N^E$  workers. The reservation wage is no longer  $W^{HC}$ , because jobs are no longer available at that wage. Instead, the reservation wage in (2) is the wage that, combined with a high effort level, would make the worker as well off as remaining unemployed with a chance of getting a higher wage,  $W^E$ , sometime in the future. This reservation wage may be above or below  $W^{HC}$ , depending on the desirability of unemployment (which depends on things like the level of unemployment insurance benefits).

In contrast to the competitive equilibrium, there is now involuntary unemployment  $N - N^E$  in the sense that the workers who are unemployed would be willing to work at the prevailing wage. In the simple supply-and-demand model, firms never offer wages that differ much from the market-clearing wage  $W^{HC}$ . If wages were above that level, workers who could not find jobs would offer to work at less than the going wage, bidding down the wage. In the efficiency wage equilibrium, workers without jobs cannot successfully underbid their employed neighbors. Recall that efficiency wages are chosen so that (2) becomes an equality. Suppose an unemployed worker approaches a firm's manager, offering to work for less than the efficiency wage. The

manager would like to pay the lower wage. But the manager understands that workers who are paid less than the efficiency wage will find it optimal to shirk, since lowering the wage makes the right-hand side of (2) less than the left-hand side. The unemployed worker's offer is therefore declined. Since all firms behave in this same manner, unemployment persists in equilibrium. (When firms differ, the result can be dual labor markets, which we discuss shortly.)

It is not hard to see how performance bonds would circumvent this problem and eliminate the involuntary unemployment. Suppose the unemployed worker approaches a firm, offering to work at less than the efficiency wage *and* offering to post a bond to be forfeited if he is detected shirking. A clever manager would understand that a big enough bond would deter shirking. The manager would accept this offer.

This last point leads to two additional observations. First, firms that pay efficiency wages will, whenever possible, want to also use partial performance bonding. Worker bonds complement efficiency wages in coaxing high effort from workers, thus reducing the efficiency-wage premium. Second, efficiency wages and resulting unemployment persist only to the extent that firms cannot resolve the agency problem by using performance bonds. If bonding schemes were costless to implement, wages would be bid down to the competitive level and unemployment would disappear.

### *Efficiency Wages and Dual Labor Markets*

The distinctive feature of efficiency-wage jobs is that they are valuable from the start; they are jobs that people want but can't easily get. The "carrot" that elicits high effort in an efficiency-wage job is the credible promise of high wages extending into the future. Efficiency-wage jobs also tend to offer stable employment. In addition, firms that pay efficiency wages might complement the efficiency-wage policy with performance bonds, so these jobs would have job ladders and pensions.

A casual look at jobs in the economy

suggests that there are some highly paid, stable jobs in which employees do work that is complicated and hard to measure. Many other jobs are characterized by simple work, poor pay, no job security, and little prospect of promotion. In short, as Doeringer and Piore (1971) argue, the American labor market seems to have a dual labor market with a "primary sector" of good jobs and a "secondary sector" of less desirable jobs. Dual labor market theorists like Doeringer and Piore argue that even hard-working, well-qualified workers in the secondary sector often cannot find employment in the primary sector. In a dual labor market, good workers can be stuck in bad jobs.

In the basic supply-and-demand model, workers with equal ability and training who are doing equally difficult or distasteful work are paid the same. In this model, there may well be poorly paid jobs, but these jobs tend to have low-skill workers doing easy work. The supply-and-demand model predicts that equally productive workers will have similar lifetime earnings. The central idea of dual labor market theory—that good workers can be stuck in bad jobs—just doesn't make sense in the competitive model. Pure performance-bonding models also envision a perfectly competitive environment, so this observation applies there, too.

Bulow and Summers (1986) argue that efficiency-wage models like the one we present here can provide an explanation for dual labor markets. Imagine a labor market in which all workers are identical but their jobs differ. In some jobs, low effort is acceptable or worker performance is easy to evaluate, so firms can effectively pay piece rates. Workers in this "secondary sector" receive a competitively determined wage. "Primary sector" jobs, in contrast, have agency problems that firms can resolve only by paying efficiency wages. All workers would like to have one of the valuable primary-sector jobs, but many well-qualified workers will end up in secondary-sector jobs.

Critics of dual labor market theories argue that labor markets efficiently sort workers into appropriate jobs, given their



ability, training, and inclinations. They argue that labor markets do not really produce primary and secondary sectors. Instead, markets sort workers according to characteristics that are not observable to labor economists (like willingness to work hard or cooperate with coworkers), creating the illusion of dual labor markets.

Critics of efficiency-wage models also point out that if there really is a secondary sector, efficiency-wage models would not imply unemployment. Instead, people who could not get high-wage jobs would accept low-wage ones. In fact, this outcome depends on how the job search is modeled. If workers cannot search efficiently for primary-sector jobs while they are employed, the equilibrium level of unemployment will make workers indifferent between searching for a high-wage job while unemployed and accepting a low-wage job. This might still be interpreted as involuntary unemployment.

### *Other Efficiency-Wage Models*

In the efficiency-wage model we outlined above, firms get higher productivity (less shirking) by paying workers more than their reservation wage. As we have seen, the market consequence of this employment-relations strategy can be dramatic. Most striking is the result that firms will not cut wages in response to involuntary unemployment, because cutting wages would reduce productivity. The effort-regulation problem we described is only one of a number of agency problems that have been addressed with efficiency-wage models. The following hypotheses about how efficiency wages might arise differ from the widely used effort-supply model in using only carrots and no sticks; the firm does not use dismissals.

**Controlling Turnover.** For many firms, orienting and training new employees can be an expensive, time-consuming activity. It can take months or even years for workers to become fully adjusted and productive in some work environments. Since firms face a big loss when employees join a firm only to quit a short time later,

reducing labor turnover is an important objective for managers. How does this problem affect compensation policy?

An employee just starting out with a firm typically won't know very much about nonwage features of the job. How difficult will the work be? Is it interesting? Are the working conditions pleasant? Will he like his boss and colleagues? Once he has spent time on the job, a typical worker will learn about these aspects of the job, and what he learns will affect his inclination to stay with the firm or seek employment elsewhere (while still employed). Indeed, the decision to quit or stay hinges on the value of the job (which in turn depends on both wage and non-wage features of the job) compared with the value of the alternative.

One option for the firm is a low-wage, high-turnover strategy. The firm can simply set the wage at the lowest level necessary to fill vacancies immediately, fully understanding that many workers will quit as they discover undesirable nonwage aspects of the job. For firms with high turnover costs, though, a better strategy will be to reduce turnover by paying a wage higher than necessary to fill open jobs.

As in the effort-regulation model, employers pay workers more than their reservation wages in order to alter their behavior. In the labor-turnover model, higher wages reduce recruiting and training costs and generate a more experienced labor force.

Salop (1979) establishes that when all firms use this strategy, involuntary unemployment can persist in the economy. Also, if firms' turnover costs differ, the market generates wage dispersion in which workers of equal ability receive different wages.

**Attracting Good Workers.** Adverse-selection models (Weiss, 1980, 1990) are based on another real-world problem that firms frequently encounter. A manager hiring a new worker wants to know how smart, conscientious, congenial, and motivated—in short, how productive—the worker is. The manager understands that workers have differing levels of productivity but can make only an informed guess

about the applicant's productivity. Often firms learn about workers' productivity only after the workers have been on the job for some time. In an extreme case, in which a firm can discern nothing about the future productivity of workers, the firm would have to resort to hiring at random from the pool of applicants.

Now suppose that, in general, the most productive workers also have the best opportunities (as self-employed workers or employees in other firms), so that more productive workers have higher reservation wages. Then if a firm offers the lowest wage necessary to fill open positions, it will be choosing from among applicants with generally low productivity. As the firm increases the wage it offers, the pool of applicants expands to include better applicants, and the average productivity of the pool increases. The firm's optimal strategy entails trading off higher wages against increased average productivity.

**Wage Norms.** The models we have discussed so far are based on the general premise that workers act in their own narrowly defined interest. Akerlof (1982) set out a "sociological" perspective on worker behavior in which the employment relationship is viewed as a "gift exchange." A firm that pays workers only the lowest wage necessary to get them to show up for work finds that workers reciprocate with minimal effort. A firm that gives workers a "gift" of higher wages (without *requiring* higher effort) finds that workers reciprocate with a "gift" of higher effort norms (which are enforced, in part, by peers). The model has characteristics similar to those of the basic effort-regulation model, but with behavioral foundations more similar to those hypothesized by sociologists than to the opportunistic utility maximization favored by economists.

Annable (1988) advances a subtle argument about the formation and rigidity of wage norms, starting from the premise that "it is a tenet of personnel management that violations of established wage relationships will lead to worker dissatisfaction." The wage relationships are both intertemporal and interpersonal and are established

either spontaneously through "equity, custom, and tradition" or by explicit coordination activity among workers. The norms thus established translate into a relationship between wages and effort (broadly defined) that the firm will find difficult to influence. The firm must therefore take this relationship as given if it chooses the profit-maximizing wage, just as in the simple effort-regulation model. Annable argues that once a set of norms has been established, they will tend to be rigid because they are a public good for workers; the benefits of the coordination activity needed to change norms are shared by all workers, not just those bearing the cost of coordination.

**Avoiding Unionization.** Union organizing entails different costs and benefits for workers and firms. The idea behind union-threat models is that by voluntarily giving workers one of the biggest benefits of unionization—higher wages—the firm can change the workers' cost-benefit calculus. Workers would still bear the cost of unionization, but the marginal benefit would be lower. If the nonwage costs of unionization (less flexible employment policies, for example) are much higher for firms than the corresponding benefits to workers, the firm would find it worthwhile to follow this approach. Of course, the firm must also believe that there is a significant chance that a union will be successfully organized if they do not act. In the right circumstances (not in the middle of an open unionization effort, for example), the firm's voluntary action could also be interpreted in Akerlof's gift exchange framework. Workers, receiving the "gift" of higher wages, believe their employer is "fair" and see no need for a union.

## EMPIRICAL STUDIES

Economic theory is most compelling when it provides plausible predictions of widespread phenomena, such as mandatory retirement, that are otherwise difficult to explain. In this section, however, we sample some of the more detailed

(but often ambiguous) empirical evidence that bears on these theories.

The simple competitive supply-and-demand model implies that wages depend only on workers' productivity and on attributes of firms or jobs that make the job more or less desirable. Characteristics such as the firm's size or the ease of monitoring employees should not affect compensation. Suppose that a worker at firm A is paid less than a worker with comparable experience, skills, and so forth at firm B. In the supply-and-demand model, the firm A worker will go to firm B and offer to work for slightly less than the current firm B worker. In the competitive supply-and-demand paradigm, then, the law of one price holds, because workers arbitrage away price differences. This observation forms the basis for most econometric tests of the different compensation models.

The performance-bonding models predict some additional relationships between wages and characteristics of workers and firms. Lazear's work-life incentives model, for instance, predicts a positive relationship between wages and job tenure (length of time in present job) after controlling for overall work experience (as well as characteristics such as education-related worker productivity). The evidence on this relationship is supportive on balance, but it is somewhat muddled by technical econometric issues (Hutchens, 1989).

Lazear's theory also predicts that delayed-payment arrangements and collateral phenomena such as mandatory retirement will not be present when employees are easily monitored (a characteristic of the job, not the employee). Hutchens (1987) bases a test on the assumption that jobs involving repetitive tasks are, on average, more easily monitored and should therefore be characterized by absence of high wages for more senior workers, mandatory retirement, pensions, and long job tenures. Despite the fact that his measure of repetitive tasks is a very noisy proxy for ease of monitoring, Hutchens finds in the National Longitudinal Survey that jobs with more repetitive tasks

are significantly less likely to exhibit the characteristics predicted by Lazear's theory.

Henry Ford is famous for deciding in 1914 to pay a wage well above the going rate. Raff and Summers (1987), who studied this episode intensively, say that "On balance it seems fair to conclude that Ford was able, by offering the five-dollar day, to reduce the turnover among his workers and to extract much more intensive, and generally productive, effort from them." Ford's policy thus had the main hallmarks of an efficiency wage: desirable effects on workers' behavior brought about by wages above the level necessary to fill vacancies.

A study by Krueger and Summers (1988) is one of a number that examine wage differentials across industries. The principle here is that, by and large, the industry in which a worker finds himself should not affect his wages in a competitive model. This observation applies to both the simple supply-and-demand model and the more sophisticated performance-bonding models (as long as average age of employees does not differ across industries). They argue that systematically higher wages for workers in one industry than in another constitute evidence of efficiency wages. Krueger and Summers show that there are significant wage differentials across industries and use various types of data to argue that these cannot be attributed to employee demographics, human capital differences, compensating differentials, or unions. Although the existence of inter-industry wage differentials is not direct evidence of efficiency wages, Krueger and Summers seem to take the position that, after all other reasonable explanations have been ruled out, the only possibility left is efficiency wages.<sup>19</sup> Murphy and Topel (1990) point out that a fully convincing explanation of inter-industry wage differentials would link wages to features of industries that, according to efficiency-wage models, should generate different wages.

Similar arguments have been made about the so-called employer-size effect; larger employers, on average, pay higher

<sup>19</sup> Gibbons and Katz (1992) give a bit more equivocal reading of the evidence on inter-industry differentials. Thaler (1989) gives a concise overview.

wages, which prove difficult to explain without efficiency wages. Rebitzer and Taylor (1995) pose a challenge to this line of reasoning. They study law firms—organizations in which there are obvious and dramatic promotion tournaments. Associates who are promoted to partner get very large increases in income, creating the presumption that the performance bonds created by the tournament are sufficient to generate high levels of effort, low quit rates, and so on. Rebitzer and Taylor show that the employer-size effect persists even in this environment, where the most common reasons for efficiency wages appear to be absent.

Cappelli and Chauvin (1991) test one component of the efficiency-wage model. Using data from a single multi-plant automobile manufacturer, they test directly whether wage premiums result in lower levels of disciplinary action. All workers in their data were covered under the same collective bargaining agreement and the same disciplinary policies. By comparing the wages specified in the contract (the same for all plants) with the average hourly wage for production work in each plant's Standard Metropolitan Statistical Area, Cappelli and Chauvin measure the wage premium paid at each plant. The premiums varied from 0 to 100 percent. They find fewer shirking-related disciplinary actions at plants with higher wage premiums. Their results provide support for a connection between pay and productivity. Because the firm was unionized, the existence of wage premiums does not imply the presence of efficiency wages, but the result does suggest that a union wage premium, by making the job valuable, acts as an efficiency wage. Of course a union contract that also makes disciplinary actions more difficult would offset that effect.

Krueger (1991) compares compensation at company-owned and franchise-owned fast-food restaurants. This comparison controls automatically for different characteristics of workers and jobs. The two groups differ because managers of company-owned restaurants have less incentive

to monitor employees than do owner-managers. Thus the two groups can be presumed to have systematically different levels of monitoring (that is,  $d$  is higher for owner-operated restaurants). Krueger finds a small wage premium and steeper tenure-earnings profiles at company-owned outlets, results consistent with the efficiency-wage model. The steeper profile would also be implied by Lazear's work-life incentives model, but the premium implies that the present value of lifetime wages is higher at company-owned outlets, for which Lazear's model offers no rationale. Interestingly, the wage premiums are much higher for low-level managers than for regular workers. This finding suggests that the incentive problems faced in this industry are most efficiently solved by paying efficiency wages to supervisors to encourage more effective monitoring of production workers.

On the other hand, using data on wages for narrowly defined occupations at 200 plants, Leonard (1987) finds that differing intensity of supervision across plants does not lead to the wage variation predicted by efficiency-wage models. We find this evidence less compelling than Krueger's because the reason for variation in supervision intensity is unobserved. Without that information, it is difficult to know whether other relevant factors are really being held constant.

## CONCLUSION

Employers and employees are often inclined to pursue goals that are at cross-purposes. The focus of this article is on economists' hypotheses about how firms resolve this problem, and on the implications of these solutions for the structure of labor markets.

Piece rates or incentive pay plans provide powerful direct incentives but have limited applicability. The performance-bonding concept adds a valuable general perspective on employment practices such as job ladders, promotion tournaments, mandatory retirement, and pension policy.

These models form an important link between labor economics and the study of firm organization. Still, there are numerous legal, institutional, and economic impediments to the use of performance bonds, so it seems likely that firms' best efforts to use this approach to motivating employees often fall short of completely resolving fundamental agency problems. Thus, even though efficiency wages are a second-best solution, they may often be needed as a complementary incentive device. Further, efficiency-wage theories present possible explanations for a number of additional labor market features, most notably involuntary unemployment.

## REFERENCES

- Akerlof, George A. "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics* (November 1982), pp. 543-69.
- \_\_\_\_\_ and Lawrence F. Katz. "Workers' Trust Funds and the Logic of Wage Profiles," *Quarterly Journal of Economics* (August 1989), pp. 525-36.
- Annable, James. "Another Auctioneer is Missing," *Journal of Macroeconomics* (Winter 1988), pp. 1-26.
- Baker, George, Robert Gibbons, and Kevin J. Murphy. "Subjective Performance Measures and Optimal Incentive Contracts," *Quarterly Journal of Economics* (1994), pp. 1125-56.
- Bulow, Jeremy I., and Lawrence H. Summers. "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment," *Journal of Labor Economics* (October 1986), pp. 376-414.
- Cappelli, Peter, and Keith Chauvin. "An Interplant Test of the Efficiency Wage Hypothesis," *Quarterly Journal of Economics* (August 1991), pp. 769-87.
- Carmichael, H. Lorne. "Self-Enforcing Contracts, Shirking, and Life Cycle Incentives," *Journal of Economic Perspectives* (Fall 1989), pp. 65-83.
- Dickens, William T., Lawrence F. Katz, Kevin Lang, and Lawrence H. Summers. "Employee Crime and the Monitoring Puzzle," *Journal of Labor Economics* (July 1989), pp. 331-47.
- Doeringer, P. B., and M. J. Piore. *Internal Labor Markets and Manpower Analysis*, Heath, 1991.
- Gibbons, Robert. "Incentives and Careers in Organizations," National Bureau of Economic Research Working Paper 5705, August 1996.
- \_\_\_\_\_ and Lawrence F. Katz. "Does Unmeasured Ability Explain Inter-industry Wage Differentials?" *Review of Economic Studies* (July 1992), pp. 515-35.
- Hart, Oliver. *Firms, Contracts, and Financial Structure*, Clarendon Press, 1995.
- Holmstrom, Bengt, and Paul Milgrom. "The Firm as an Incentive System," *The American Economic Review* (September 1994), pp. 972-92.
- Hutchens, Robert M. "Seniority, Wages and Productivity: A Turbulent Decade," *Journal of Economic Perspectives* (Fall 1989), pp. 49-64.
- \_\_\_\_\_. "A Test of Lazear's Theory of Delayed Payment Contracts," *Journal of Labor Economics* (October 1987, part 2), pp. S153-70.
- Jensen, Michael C., and Kevin J. Murphy. "CEO Incentives—It's Not How Much You Pay, but How," *Harvard Business Review* (May-June 1990), pp. 138-49.
- Krueger, Alan B. "Ownership, Agency, and Wages: An Examination of Franchising in the Fast Food Industry," *Quarterly Journal of Economics* (February 1991), pp. 75-101.
- \_\_\_\_\_ and Lawrence H. Summers. "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* (March 1988), pp. 259-93.
- Lazear, Edward P. "Why Is There Mandatory Retirement?" *Journal of Political Economy* (December 1979), pp. 1261-84.
- \_\_\_\_\_. "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *The American Economic Review* (September 1981), pp. 606-20.
- \_\_\_\_\_. *Personnel Economics*, The MIT Press, 1995.
- \_\_\_\_\_ and Sherwin Rosen. "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* (October 1981), pp. 841-64.
- Leonard, Jonathan S. "Carrots and Sticks: Pay, Supervision, and Turnover," *Journal of Labor Economics* (October 1987), pp. S136-52.
- Malcomson, James M. "Work Incentives, Hierarchy, and Internal Labor Markets," *Journal of Political Economy* (June 1984), pp. 486-507.
- Medoff, James, and Katharine Abraham. "Experience, Performance, and Earnings," *Quarterly Journal of Economics* (December 1980), pp. 703-36.
- Murphy, Kevin M., and Robert H. Topel. "Efficiency Wages Reconsidered: Theory and Evidence," *Advances in the Theory and Measurement of Unemployment*, Yoram Weiss and Gideon Fishelson, eds., St. Martin's Press, 1990, pp. 204-40.
- Raff, Daniel M. G., and Lawrence H. Summers. "Did Henry Ford Pay Efficiency Wages?" *Journal of Labor Economics* (October 1987), pp. S57-86.
- Rebitzer, James, and Lowell J. Taylor. "Efficiency Wages and Employment Rents: The Employer Size Wage Effect in the Job Market for Lawyers," *Journal of Labor Economics* (October 1995), pp. 678-708.
- Ritter, Joseph A., and Lowell J. Taylor. "Workers as Creditors: Performance Bonds and Efficiency Wages," *The American Economic Review* (June 1994), pp. 694-704.



# REVIEW

SEPTEMBER/OCTOBER 1997

---

\_\_\_\_\_ and \_\_\_\_\_. "Seniority-Based Layoffs as an Incentive Device," Federal Reserve Bank of St. Louis Working Paper 97-17A, November 1997.

Shapiro, Carl, and Joseph Stiglitz. "Involuntary Unemployment as a Worker Discipline Device," *The American Economic Review* (June 1984), pp. 433-44.

Thaler, Richard H. "Anomalies: Inter-industry Wage Differentials," *Journal of Economic Perspectives* (Spring 1989), pp. 181-93.

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

\_\_\_\_\_. *Efficiency Wages: Models of Unemployment, Layoffs, and Wage Dispersion*, Princeton University Press, 1990.