2
Fixed or Variable Pay?

Variable pay simply means tying a worker's compensation to some output-based measure of performance. Fixed pay means that the worker's compensation is independent of output, often because output is difficult to measure or define, or because variations in output are affected primarily by factors over which the worker has no control.

In the real world there are many forms of variable pay. Indeed it is possible to argue that almost no job's pay is truly fixed. Poor performance over a substantial period of time results in lower future pay and, at the extreme, termination. Let us begin with the most basic form of variable pay, namely the piece rate.

Piece Rate Pay

Before asking whether the firm should compensate its workers according to a piece rate scheme, let us first derive the optimal scheme for a worker who is risk neutral.[1] Optimal schemes must accomplish two things, that is, they must achieve efficiency on two margins. First, they must induce a given worker to put forth the appropriate level of effort. Second, they must induce the right workers to come to work for the firm. Under risk neutrality the optimal payment scheme is a linear one, achieving the first-best level of effort.

To see this, consider the problem of a firm that is attempting to decide on the commission rate to give to a salesperson. The firm wants to maximize profits, but it must pay the worker enough to induce him to work for the firm. The problem can be broken into two stages. The first stage is labor supply: It is necessary to determine what level of effort or number of hours of work a worker will supply given some structure of compensation. Then, given the worker's labor supply behavior, the firm must choose the compensation formula that maximizes profits.

Let the firm pay on the basis of output according to the scheme

$$\text{Pay} = \alpha + \beta q, \tag{2.1}$$

where q is output and α and β are compensation parameters to be chosen by the firm. Output depends on effort and luck. Normalize the measurement of effort so that one unit of effort produces one unit of output. Then

$$q = e + v,$$

where v measures luck or measurement error.

The worker likes income but hates work, and his distaste for work is given by $C(e)$, where both C' and C'' are positive. The conditions C', C'' 0 guarantee that the solution implies finite effort levels. These are not assumptions but rather empirical dicta. At some point the cost of producing an additional unit of expected output becomes infinite as the worker reaches complete exhaustion.

The worker's labor supply function is the solution to

$$\max_{e} E[\alpha + \beta(e + v)] - C(e), \tag{2.2}$$

with first-order condition

$$C'(e) = \beta. \tag{2.3}$$

Equation (2.3) is the worker's labor supply function, which the firm takes as given when it maximizes profits by choosing the parameters α and β. Given C'' 0 and risk neutrality, effort increases in β. Higher "wage rates" induce more effort or hours of work, so labor supply functions are positively sloped. The firm's problem is

$$\max_{\alpha,\beta} E(q) - (\alpha + \beta e),$$

$$\text{or} \tag{2.4}$$

$$\max_{\alpha,\beta} e - (\alpha + \beta e)$$

subject to the individual rationality constraint that the worker is willing to take the job in the first place. Required is that

$$\alpha + \beta e \geq C(e). \tag{2.5}$$

Equation (2.5) merely says that the worker must earn enough to cover his disutility at the equilibrium level of effort. Substitution of (2.5) into (2.4) yields

$$\max_{\alpha,\beta} e - C(e), \tag{2.6}$$

with the first-order condition

$$\frac{\partial}{\partial\beta} = [1 - C'(e)]\frac{\partial e}{\partial\beta} = 0. \tag{2.7}$$

($de/d\alpha = 0$, so the second condition is redundant.)

Equation (2.7) implies that the firm will choose β so as to bring about efficiency. The firm, in its quest for profit, induces the worker to set the marginal cost of effort equal to its marginal social value of effort, in this case 1. Equations (2.3) and (2.7) taken together imply that $\beta = 1$. After β is chosen, the optimum level of effort is determined by (2.3); (2.5) then dictates the size of α necessary to attract the worker to the firm.

The fact that $\beta = 1$ implies that piece rate workers should be entitled to the entire residual profit. The firm merely "rents" the worker the job at price α and then gives the worker the full output. The value to paying a worker 100 percent of the residual is illustrated by a taxicab example. There are many ways to pay taxi drivers. One possibility is to rent them the cab and allow them to keep everything that they make for themselves after having paid the rental. This is the scheme that has just been shown to be optimal. An alternative is to give them the cab and take 50 percent of the meter. The second scheme has a couple of problems. Among the most obvious is that since the company cannot monitor the amount of driving done, the cabby and passenger can make a deal to pay 75 percent of what the meter would show and make the trip with the meter off. Both passenger and cabby are better off at the expense of the company. When the meter is off, the cab's "for hire" sign is lighted so that the company can detect cheating. Still this is costly, and it provides a rationale for making the driver the complete residual claimant.

Another rationale, and the one on which I focus now, is that a 50 percent commission rate induces insufficient effort by the driver. Consider a taxi driver who has been working for eleven hours in a particular day and is trying to decide to whether to drive the cab for a twelfth hour. He reasons that if he drives the cab, he expects to pick up $10 in cab fare that hour. Let's suppose that he values the leisure associated with that twelfth hour of work at $8. That is, at any price greater than $8, the taxicab driver would be willing to put forth the effort and drive the cab. It is clearly efficient for the driver to work during that twelfth hour. The value of driving the twelfth hour is measured by the price that people are willing to pay for the services (in this case $10). The social cost associated with driving that twelfth hour is the driver's reservation wage, in this case \$8. Since 10 exceeds 8, the driver should drive the cab. Put alternatively, both the cab company and the driver can be made better off by having the driver work the twelfth hour. Since the cab company takes in, at most, \$10 in revenue and the driver requires, at least, \$8 of compensation, deals can be struck that will make both parties happy and will induce the cab driver to drive the cab. But splitting the revenues is not one of them.

Suppose that the cab driver takes home 50 percent of his compensation and does not pay any rental on the cab. In this case the driver will quit and will not supply the twelfth hour of effort. Since he takes home only half of the \$10 fare and since his reservation wage is \$8, he prefers the leisure to the \$5 of revenue. This is the standard intuition behind the rule that says a worker must be made a full residual claimant in order to induce him to put forth the efficient amount of effort.

There is another reason to make the worker receive the full marginal output, which has nothing to do with effort.[2] If workers differ in their characteristics, then paying a worker less than the full amount will cause adverse selection problems in the hiring process. The wrong workers will want to work at the firm.

Suppose that workers have different ability levels, indexed by q, which signify the level of output that they can produce. Let one firm pay according to the optimal formula derived above:

$$\text{Pay} = a + q.$$

Let another pay at a rate $b < 1$ for every dollar of output, but let it offer a (potentially) higher base wage, a'.[3] The situation is shown in figure 2.1.

All those workers with $q < q^*$ prefer the low piece rate firm, namely the one that sets $b < 1$ but sets $a' > a$. All those with $q < q^*$ prefer the high piece rate firm. Thus the best (and most profitable) workers go to the high piece rate firm, while the low-quality workers choose the firm with a higher base salary and lower commission rate.

It is necessarily the case that the firm makes more money on workers whose ability exceeds q^* than those whose ability is less than q^* when the compensation schemes are given by $a' + bq$ for low-ability workers and $a + q$ for high-ability workers. As long as there is free entry into the industry, firms cannot receive any rents ex post. In equilibrium, firms that pay $a + q$ will have to set a equal to minus the rental price on capital. The $|a|$ cannot exceed the rental price of capital, or workers could be stolen away by other firms that charge only a for the capital. Now consider firms that are charging only a' for the job. The worker's output, net of

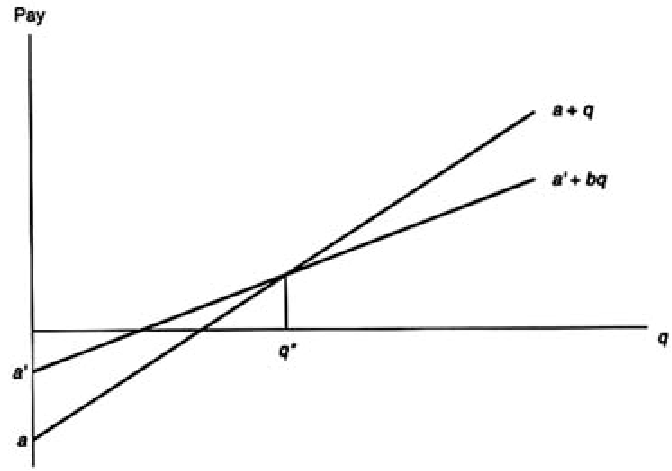

Figure 2.1

capital cost, is measured by $a + q$, but the worker is receiving $a' + bq$. In the region where q is less than q^*, $a' + bq$ exceeds $a + q$, which means that the worker is being paid more than his net output. Thus all firms are losing money on workers up to q^*. For workers whose output exceeds q^*, firms that paid $a' + bq$ would do better if they were able to attract the workers. Unfortunately, if some firms offer $a + q$ (which in equilibrium they must), all high-ability workers will prefer the firm that pays $a + q$ because that is where those workers receive the highest earnings. Thus workers who have $q < q^*$ choose to work at firms that pay $a' + bq$, but those firms lose money. All workers who have ability levels greater than q^* choose the firms that pay $a + q$, and those firms break even on the workers. Since breaking even is better than losing, profits are higher for firms that choose the $a + q$ strategy.

As mentioned, many taxicab companies make their drivers full residual claimants after renting the driver the cab. At first glance this arrangement does not seem typical of the standard salesperson's contract. Few salespeople receive 100 percent commission rates. It is more common for firms to pay, say, 10 percent of sales revenue to their salesperson.

The evidence on this point is less clear than it appears. Recall that the worker should be made full residual claimant and that β should be set equal to one. But β multiplies net output, not sales. A residual claimant receives revenues after other variable factors of production have been paid. In the

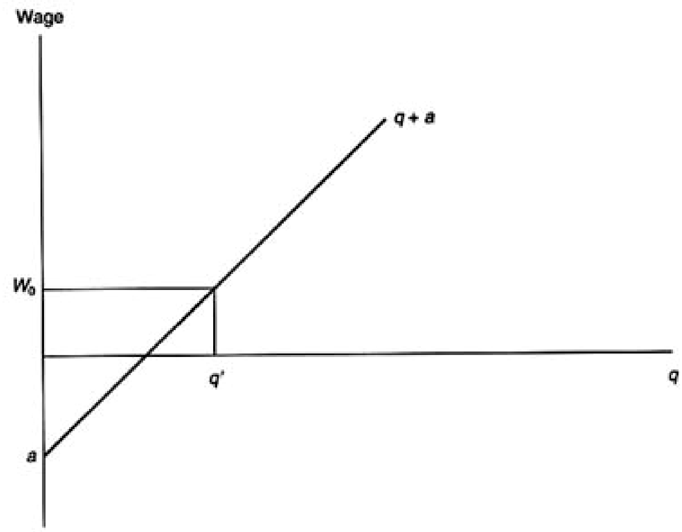

Figure 2.2

case of a computer salesperson, workers who produce the computers and material suppliers must be paid out of revenues. If other costs of production accounted for, say, 80 percent of total revenue, then paying a 20 percent commission rate on sales would be equivalent to paying a 100 percent commission rate on profit as the theory prescribes. So things may not be as far from the risk-neutral model as they first appear.

Another seeming departure from reality is that salespeople generally do not start out with negative earnings. On the contrary, they often receive "draws" or advances against their commissions. Still this is almost identical to the scheme already described where the vertical intercept a is negative. In figure 2.2 the worker receives a draw of W_0, which means that no commissions are paid until q' units are sold. Then the worker receives the standard commission on all units greater than q'. This kinky compensation scheme is nothing more than a spline of a horizontal segment with the line $q + a$. The horizontal segment is only relevant if q falls short of q'. However, in reality, if q falls short of q' repeatedly, the worker loses his job. Thus compensation is dominated by the part of the payment function that is identical to $q + a$.

The amount that a worker is willing to pay for the job (the magnitude of α) depends on the amount of capital with which he works. Drawing on a real world example, Elton (1991) has examined the compensation of American stockbrokers. He found that in large firms with good reputations, the implicit α is more negative than in firms with lesser reputations. This is as it should be. It is easier to sell stock when working for Merrill Lynch than when working for Lazear Securities. Merrill Lynch sets a lower α, consistent with a higher q' for continued employment. It is the constant term, not the slope parameter, that varies across firms.

The firm's CEO has control over the most capital. For this reason standards (the implicit q') are very high. In the case where the CEO is paid $a + q$, the implied value of a is the amount of capital that the CEO rents. In this case it is the entire value of the firm. But it would be impossible to extract such a large payment from an individual. Of course, in equilibrium, no payments must be extracted. The CEO earns a positive salary after having paid the rent, not a negative one. Rather than specifying the formula as explicitly containing a negative intercept term, boards of directors instead give the CEO a fixed wage w_0, as shown in figure 2.2, and set an extremely high q' below which the CEO does not receive any bonus compensation. As long as output is always greater than q', the schemes are equivalent. If output falls below q', the CEO will find that his job is in jeopardy. The implicit q' below which output may not fall without termination is much higher for CEOs than for other workers. As such, the highest-ability workers are selected to be CEOs.

The taxicab example makes clear that setting $\beta = 1$ does not solve all problems. Even if the driver rents the cab and keeps all revenue collected, he still does not have the appropriate incentives to care for capital entrusted to him. Rented cabs tend to be in worse shape than owner-operated cabs. Some have argued that in American football the ability of players to negotiate on their own behalf, independent of their team's interest, has resulted in more injuries. Since team owners cannot capture the full return on the player, teams have less incentive to protect the players' health.[4]

Payment by Input versus Payment by Output

Workers are sometimes paid on the basis of some time unit, such as an hour, day, week, or year, and sometimes on the basis of some measured output, such as the number of pieces of fruit that they pick. How do firms choose between the various compensation methods? If they decide, say, to pay on the basis of output, what formula is used? Salesperson's commission pay, for example, can take many forms and shapes. Which is the