Milgrom, Roberts

is zero. Efficient risk sharing requires shifting all the risk onto the risk-neutral party, who suffers no cost in bearing the risk.

This conclusion, however, depends on ignoring the incentive problems for insurance and employment created by the condition of moral hazard.

Principles of Incentive Pay

The general problem of motivating one person or organization to act on behalf of another is known among economists as the *principal-agent problem*. This problem encompasses not only the design of incentive pay but also issues in job design and the design of institutions to gather information, protect investments, allocate decision and ownership rights, and so on. However, we focus our discussion in this chapter principally on the issues surrounding incentive pay, and we set our discussion of incentives in the context of employment. The principal in this case is the employer, who wants the employee (the agent) to act on his or her behalf.

Basing Pay on Measured Performance

As we discussed in the introduction to this chapter, there are many situations in which providing incentives requires that employees' pay depend on their performance. Essentially, if the employees' direct provision of effort, intelligence, honesty, and imagination cannot be easily measured, then pay cannot be based on these and any financial incentives must come from basing compensation on performance. Efficient risk sharing, in contrast, requires that each person in society should bear only a tiny share of each risk, without regard to its source. In particular, individuals should be insulated against the randomness that would enter their pay by basing it on measured performance. Therefore, performance-based compensation systems cause a loss from inefficient risk sharing. The money value of the loss is equal to the risk premium associated with the actual compensation system minus the risk premium that would be associated with efficient risk sharing. Firms that use performance-based compensation hope to recoup this loss (and more) by eliciting better performance from their employees.

There are various reasons why incentives might be needed to elicit top-notch performance. Some employees may find their work distasteful and may neglect it unless they are held responsible for achieving results. Even when employees are hard workers who like their jobs, they may still have priorities that are different from those of their employer. For example, without compensating incentives, managers might be tempted to be too generous to their subordinates in granting raises and time off, or to hire the children of relatives and friends, to spend lavishly on a pleasant work environment or on fancy accommodations when traveling on business, to use company resources for community projects that raise their personal status, to devote excessive efforts to projects that advance their careers or that are especially interesting or pleasant, and so on.

To analyze these possibilities in a model, we suppose that the employee must exert an effort e at personal cost $C(e)$ to serve the interests of the employer. The effort e represents any activity that the employee undertakes on behalf of the firm, and the cost $C(e)$ can represent the unpleasantness of the task, foregone perquisites, lost status in the community, or anything else that the employee gives up to serve the employer's interests. For tasks that are pleasant, the "cost" can be zero or even negative.

The effort e is assumed to determine to the firm's profits: Profit $= P(e)$. It is sensible to assume that greater effort leads to higher profits. It is not necessary for the employer actually to know the functional relationship between effort and results; instead, the P function can be thought of as the employer's *subjective* estimate of the productivity relationship. If the relationship between profits and effort is random, then $P(e)$ should be thought of as the expected value of profits when effort level e is expended.

It may be impossible for anyone to observe an employee's direct effect on profits, but it is that effect, in principle, that the employer cares about. For example, the employee may be a sales representative whose efforts lead to no sales today but create a good impression that brings customers back in the future. The employer may care about the impression that is created, without actually being able to tell either how hard and how skillfully the employee has tried to impress customers or how many customers have actually been favorably impressed.

The general point here is that compensation can vary systematically only with things that the employer can observe. The employer cannot pay more to sales representatives who are particularly effective in creating a good impression if it is impossible to tell who they are. In addition, even some observable indicators may not be suitable bases for compensation. It may be possible, in principle, for the manager to photograph the faces of customers as they leave the store and pay compensation based on how many faces were smiling. What makes this possibility seem so absurd is its manifestly subjective nature. What is a "smiling" face? To base a compensation formula on something that is not objectively measurable is to invite disputes and unhappiness among employees.

A Model of Incentive Compensation

For our first formal model of incentive compensation, we assume that the effort level e that the employee chooses can be understood to be a number—for example, energy expended or hours worked. As we have already noted, if e were directly observed, there would be no difficulty in providing adequate incentives; the employer could make pay contingent on satisfactory performance without exposing the employee to any risk. We therefore suppose that the effort e cannot be directly observed. We shall suppose, however, that the employer can observe some imperfect indicators of e, that is, indicators that provide some information about e but are contaminated by random events beyond the control of the agent. For example, measured output might provide such a signal: It is related to effort, but many influences beyond the employee's control also affect the realized output. In addition, the employer may be able to observe other indicators of factors, such as general economic conditions, that are not controlled by the employee but that do affect performance.

Suppose that the indicator of effort can be written in the form $z = e + x$, where x is a random variable, and that a second indicator is y, where y is not affected by the effort e but may be statistically related to x, the noise between e and the observed z. Note that e and x are not separately observed; only their sum, z, is observed, and many different combinations of e and x yield the same level of observed z. Thus, high effort might be offset by bad luck, or low effort might be masked by good fortune.

For example, if the employee is the sales manager for some product, z might be a measure of total sales for the product (which depends on sales effort, e, and random events, x, such as realized demands) and y might measure total industry demand, which is correlated with the potential demand in the markets where the employee manages sales and thus with realized sales. To keep our formulas as simple as possible, we suppose that x and y are each adjusted to have mean zero. Then, the expected level of sales is just the effort level. In terms of the example, instead of making y the industry demand, we could make it the amount by which industry demand differs from a forecast value.

The class of compensation rules that we study are those that are linear in the two observations, that is, ones that can be written in the following form, where w stands for wage:

$$w = \alpha + \beta(e + x + \gamma y) \tag{7.3}$$

Compensation thus consists of a base amount, α, plus a portion that varies with the observed elements, z and y. We use β to measure the **intensity of the incentives** provided to the employee, so that one contract will be said to provide "stronger incentives" than another if the first contract specifies a higher value for β. The justification for this language is that if the employee increases his or her effort choice e by one unit, then according to Equation 7.3, expected compensation increases by β dollars, so higher levels of β bring greater returns to increased effort.

The parameter γ indicates how much relative weight is given to the information variable y (as compared to $z = e + x$) in determining compensation. If γ is set at zero, then y is not used in determining compensation. Given any value for γ, the term $z + \gamma y$ gives an estimate of the unobservable e. One of the principle issues in contract design is to determine how much, if any, weight to give to y in this estimate, that is, to determine the level of γ.

As an example of such a contract, suppose α is \$10,000, β is \$20 and γ is 0.5. Then expected pay is \$10,000 + \$20e, because the expected values of x and y are zero. If the employee sets e equal to 100, the expected pay becomes \$12,000 (= \$10,000 + \$2,000); if e is set at 200, the expected pay is \$14,000. Unless there is no real uncertainty, however, x and y will often not take on their expected values, and so pay will deviate randomly from its expected level. If x is more favorable than expected, say taking on the value 100, whereas y is less favorable, taking on the value -400, then the observed values are $z = e + 100$ and $y = -400$. Now an effort level of $e = 100$ brings pay of \$10,000 + \$20(100 + 100 + 0.5(−400)) = \$10,000, and an effort level of 200 brings pay of \$12,000. Of course, if x and y take on different values than those just specified, the compensation again will differ. For example, with $e = 100$, $x = -100$ and $y = 100$, pay is \$11,000, whereas effort of 200 with these same levels for the random factors brings an income of \$12,000. Thus, pay varies not just with the employee's effort, but also with the random events represented by x and y, and this randomness imposes risk on the employee (unless β is zero).

THE LOGIC OF LINEAR COMPENSATION FORMULAS The restriction to linear compensation formulas such as the one in Equation 7.3 is not always sensible. The ideal form of the compensation rule in any circumstance depends on the nature of the efforts required and on the available performance measures. Linear compensation formulas are quite popular, however, and so we take a brief diversion from our main analysis to consider when such schemes might work especially well. The considerations that arise in this discussion should serve as a reminder that incentive compensation issues are very complicated ones and not all of the relevant issues are represented in our simple mathematical models.

Linear compensation formulas are commonly observed in the form of commissions paid to sales agents, contingency fees paid to attorneys, piece rates paid to tree planters or knitters, crop shares paid to sharecropping farmers, and so on. Linear formulas are not the only ones used, however. For example, sales agents are sometimes paid a bonus for meeting a sales target. As compared to a system of sales commissions, a reward for meeting a sales target has the disadvantage that the sales representative loses any special incentive to make additional sales after the target is reached or after a poor start leaves the target hopelessly out of reach. Commission systems apply a uniform "incentive pressure" that makes the agent want to make additional sales regardless of how things have gone in the past. In selling, because incremental sales are typically equally profitable for the firm after either a slow or a fast start, this uniform incentive pressure is appropriate (in fact, optimal).

Partly as a result of efforts by firms to avoid the problem just described, when sales targets are used they are often set to cover short periods of time, so that the periods during which incentives are too low are not extended ones. This makes the compensation of additional sales efforts more nearly equal over time. The sales representatives themselves can be expected to respond to time-varying incentives by advancing or delaying the closing of sales until the period when the compensation rate is highest. To the extent that the sales representatives succeed, they have effectively arranged for all sales to be compensated equally, that is, they have converted what is nominally a sales target system into something closely resembling a system of commissions proportional to sales.

Beyond this, of course, linear systems have the advantage of being simple to understand and administer. A scheme that employees cannot understand or that cannot be administered as intended cannot provide the desired motivation.

TOTAL WEALTH UNDER A LINEAR CONTRACT An employee's ability to bear risk is negligible compared to the employer's whenever the employer is a large or medium size enterprise. For this reason, it would be optimal—incentive issues aside—for the employer to bear all financial risks, leaving the employees fully insured against all sources of fluctuation in their incomes. However, removing all compensation risk also removes all the employee's direct financial incentives to increase profits by providing effort. What is wanted is an employment contract that balances the need for risk sharing against the need to provide incentives.

Actual employment contracts involve a large number of terms, but we wish to focus on only those few dealing directly with incentive pay. Therefore, we will characterize a contract by a list of parameters $(e, \alpha, \beta, \gamma)$ that specify what level of effort e the employer expects to elicit and how the employee is to be compensated on the basis of performance. The employee's certain equivalent wealth from such a contract is the expected compensation paid minus the personal cost to the employee of supplying effort minus any risk premium: $\alpha + \beta(e + \overline{x} + \overline{y}) - C(e) - \frac{1}{2}rVar[\alpha + \beta(e + x + \gamma y)]$, where $\overline{x}$ and $\overline{y}$ are the mean levels of x and y and r is the employee's coefficient of absolute risk aversion. Recall that, to simplify formulas, we had assumed that both $\overline{x}$ and $\overline{y}$ are zero. Using the formulas about variances in the appendix, we find that the employee's certain equivalent income consists of expected income minus the cost of effort and minus a risk premium for the income risk the employee bears:

$$\text{Employee's Certain Equivalent} = \alpha + \beta e - C(e) - \tfrac{1}{2}r\beta^2 Var(x + \gamma y). \quad (7.4)$$

The employer's certain equivalent consists of the expected gross profits minus the expected compensation paid:

$$\text{Employer's Certain Equivalent} = P(e) - (\alpha + \beta e) \quad (7.4a)$$

Implicit in this is a hypothesis that the employer is approximately risk neutral.

Notice that the employee's certain equivalent consists of α plus a function of the other variables (β, γ, e) and the employer's consists of $-\alpha$ plus another function of those variables. That is, each party's equivalent wealth consists of a money term plus a term that depends on all the other aspects of the decision. By transferring money from one party to the other, one can raise one party's certain equivalent and reduce the other's by an equal amount. This is precisely the no wealth effects condition that we described in Chapter 2; we can therefore apply the value maximization principle. It follows that any efficient contract must specify the parameters so that they maximize the sum of the certain equivalent incomes of the two parties. That sum is

$$\text{Total Certain Equivalent} = P(e) - C(e) - \tfrac{1}{2}r\beta^2 Var(x + \gamma y) \qquad (7.4b)$$

Equation 7.4b specifies what is to be maximized.

INCENTIVES FOR EFFORT AND CONTRACT FEASIBILITY The next step is to specify which choices of contracts are feasible. After all, it would be ideal to ask the employee to work hard without having to provide any incentives or make the employee bear any risk! We require, however, that the employer be realistic: The level of effort the employer expects must be compatible with the incentives that are provided to the employee. Although the anticipated effort level of the employee is part of the contract, the actual effort level cannot be directly observed and is chosen later by the employee, with his or her own interests foremost in mind. To be realistic, we (and the employer) must therefore determine how the employee's choice of effort e will depend on the other parameters (α, β, γ) of the contract.

Equation 7.4 provides the key to the answer. Suppose that the costs of providing effort vary smoothly with the level provided and that the cost of effort increases at an increasing rate or, in other words, the marginal cost of effort to the employee is rising. Then, the level of effort that maximizes the employee's certain equivalent income in Equation 7.4 is the level that makes the derivative of that expression equal to zero, that is,

$$\beta - C'(e) = 0 \qquad (7.5)$$

Equation 7.5 is called an *incentive constraint* and must be satisfied by any feasible employment contract. It says that employees will select their effort levels in such a way that in their marginal gains from more effort equal their marginal personal costs. The gain is the increased pay, and a unit increase in effort brings an expected increase in pay of β; the marginal cost is C', the rate at which the personal cost of effort increases as the level provided increases.

An employment contract is therefore efficient if and only if the choices (e, α, β, γ) are ones that maximize the total certain equivalent in Equation 7.4b among all "incentive-compatible" contracts, that is, among all contracts that are consistent with Equation 7.5 and thus realizable or feasible. It is useful to solve problems of this kind in two steps. In the first step, we fix the effort e at some level and ask how the parameters α, β, and γ are optimally chosen then. This is called the **implementation problem** of obtaining the specified level of effort in the most efficient fashion.

It is evident from Equation 7.5 that fixing e also amounts to fixing β at $C'(e)$ if we are actually going to get the employees to provide the specified effort level. In Figure 7.1, to raise the effort level that the employee will choose to provide from e to $\bar{e}$ necessitates increasing the intensity of incentives from β to $\bar{\beta}$. The difference in the intensity of incentives needed can be computed as the difference in the desired effort levels times the slope of the marginal cost-of-effort curve, C''.

Also, from Equation 7.4b, we see that α does not affect the total certain equivalent at all (it determines only how the total is divided between the two parties). Thus, putting aside any requirement that both parties be willing to agree to the contract (which would limit the possible values of α to ensure that each's expected welfare was sufficiently high), we see that the efficiency of the contract does not depend on the choice of α. As for γ, it is clear that the total certain equivalent is maximized when γ is chosen to make $Var(x + \gamma y)$, the variance of the estimate of e, as small as possible because this minimizes the risk premium—the costs of imposing risks on the employees to generate incentives.

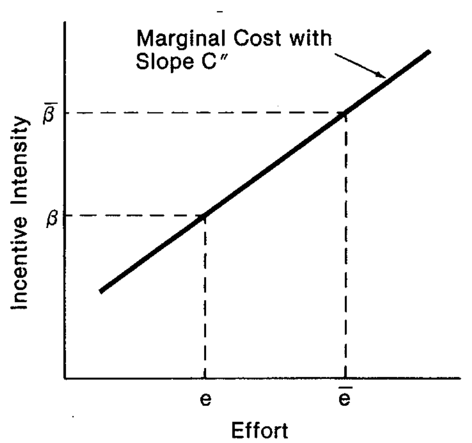

Figure 7.1: Increasing effort provided from e to $\bar{e}$ requires increasing ß to $\bar{ß}$, where $\bar{ß} - ß = (\bar{e} - e)C''$.

The Informativeness Principle

This last result—that γ should be chosen to minimize the variance of $x + \gamma y$, the estimate of e—is a special case of a more general principle.

> **The Informativeness Principle.** In designing compensation formulas, total value is always increased by factoring into the determinant of pay any performance measure that (with the appropriate weighting) allows reducing the error with which the agent's choices are estimated and by excluding performance measures that increase the error with which effort is estimated (for example, because they are solely reflective of random factors outside the agent's control).

As applied to our particular model, a measure with low error variance serves as a better basis of performance pay than a measure with higher variance. Thus, y should be included in the determinants of pay if and only if there is some value for γ that makes $Var(x + \gamma y)$ smaller than $Var(x)$, the estimate that results when y is ignored and γ is set at zero. The optimal value for γ is determined by minimizing $Var(x + \gamma y)$.

Using appendix Equation 7.18, we see that $Var(x + \gamma y)$ equals $Var(x) + \gamma^2 Var(y) + 2\gamma Cov(x, y)$, where $Cov(x, y)$, the **covariance** *of x and y*, is a statistical measure of how x and y are related and vary together. Minimizing this expression with respect to γ yields the result that γ should optimally be set at $-Cov(x, y)/Var(y)$.

If x and y are independent, then $Cov(x, y)$ is zero. In this case, γ is optimally set equal to zero. This reflects the fact that with x and y independent, knowing y tells us nothing about x and so gives us no better estimate of e: There is no point in simply adding noise to the performance measure. If x and y are positively related, as they might be if x reflects the conditions in a specific market and y is a measure of general market conditions, then $Cov(x, y)$ is positive. Then γ should be negative. Good general market conditions (positive levels of y) likely mean that conditions were also good in the specific market (positive x). Therefore, a greater portion of any given level of the observed performance $z = x + e$ is likely to reflect good luck (high x) rather than good effort (high e). Similarly, if y is low, x was also likely to be low, and a given z signals a higher level of effort e. A negative value for γ takes account of these likelihoods by increasing pay when general conditions are bad and decreasing it when they are good. Meanwhile, if x and y tend to move in opposite directions from one another, so that a low y is likely to correspond to a high x and vice versa, then $Cov(x, y)$ is negative and γ is optimally positive. A high y then signals that the given, observed level of z was likely obtained despite a low level of x, and therefore a high y is evidence suggesting a high level of e, which is rewarded through a positive value for γ.

Also note that as the variance of y increases, the magnitude of γ optimally decreases. Larger values of $Var(y)$ mean more "noise"—less reliable information—and the optimal choice of γ takes account of that by giving less weight to the signal. Even if y is an extremely unreliable measure, it will still optimally be used, but it will be given very little weight, affecting pay significantly only when it takes on an extremely large or small value.

APPLICATION: COMPARATIVE PERFORMANCE EVALUATION In applying the informativeness principle, consider the practice of **comparative performance evaluation**, according to which the compensation of an employee (typically a manager or executive) depends not just on his or her own performance but on the amount by which it exceeds or falls short of someone else's performance. Debates about this practice often revolve around the issue of controllability: As a matter of principle, it is argued, an employee's compensation should not depend on things outside the employee's control because that is perceived as unfair and because it appears to make the employee bear an unnecessary risk. So when is comparative performance evaluation a good idea? When would it be better to base the compensation of the employee only on his or her own performance?

To phrase this issue in the terms of our theory, suppose the measured performance of the employee depends on the employee's efforts, on random events that affect that employee only, and perhaps on other factors that affect all similarly situated employees. For example, the employee's measured performance might depend on the difficulty of the task, which is similar to that of the tasks assigned to other workers. Or, if the employee is a manager, the profitability of his or her unit might depend on what happens to oil prices, or interest rates, or the general level of demand in the industry. Each of these factors could be expected to have a similar effect on the profits earned by other similarly situated units.

To formalize all this, suppose there are two managers, A and B. Suppose the performance measure for manager A can be written in the form $z = e_A + x$, where e_A is the effort of manager A and x is the sum of two independent components: $x = x_A + x_C$. In this expression, x_A is a random component that affects A's performance only and x_C is a random component that affects both A's and B's performances. (The subscript C stands for this "common" source of randomness.) Similarly, B's performance measure takes the form $y = e_B + x_B + x_C$, where x_A, x_B, and x_C are independent sources of randomness. Is it better to compensate manager A based on the *absolute* performance measure $z = e_A + x_A + x_C$ or on the *relative* performance measure $z - y$, which is equal to $e_A - e_B + x_A - x_B$?

The informativeness principle directs us to the error variances attached to each compensation scheme. The variance of the first (absolute) performance measure is $Var(x_A) + Var(x_C)$, whereas the variance of the second (relative) is $Var(x_A) + Var(x_B)$ (again, see the formulas in the appendix). The relative performance measure therefore has lower variance and is to be preferred if and only if $Var(x_B) < Var(x_C)$. In other words, if the randomness that affects performance is predominantly due to a common effect, such as oil price increases or the unknown difficulty of the task, and if the variation in performance due to random events that affects particular people is smaller than the variance of the common element, then comparative performance evaluation is better than individual performance evaluation because it enables the employer to eliminate the main source of randomness in evaluating performance. If the reverse relation holds ($Var(x_C) < Var(x_B)$), however, that is, if common sources of randomness that affect both employees have smaller effects than does the randomness that affects individual employees, then it is better to base compensation on an absolute standard of performance.

Of course, in general, neither purely absolute nor purely relative performance evaluation is most efficient. As the informativeness principle establishes, some mix of absolute and comparative performance evaluation is generally preferred to either extreme form. In fact the relative weights to be placed on $e_A + x_A + x_C$ and on y can be computed from the principle.

APPLICATION: DEDUCTIBLES AND COPAYMENTS IN INSURANCE In automobile insurance, *collision* coverage is insurance that pays the owner of an automobile when his or her own auto is damaged in a collision. *Comprehensive damage* coverage is insurance that pays for damage to the person's automobile when it is stolen or damaged by other means, such as by a falling tree in a storm. Both of these kinds of coverage usually work by specifying a *deductible*, which is the portion of the loss that the insured person must pay before any payment is due from the insurance company.

Suppose that the owner of the car can, by driving carefully, parking in a garage, keeping the car doors locked, and so on, reduce the probability that the car will be stolen or damaged. That is the kind of effort that the insurance company would want to elicit. In the case of a collision or a theft, however, the owner has no control over the size of the loss that would be suffered. In that case, the size of the loss provides no information about the care taken by the owner. Therefore, according to the informativeness principle, the owner's contribution toward any loss should not depend on the size of the loss but only on the most informative performance indicator, which is the fact that a loss has occurred. So, in an optimal insurance contract, the owner's contribution should not depend on the size of the loss but rather should be a fixed amount per accident, which is very nearly the terms of a standard auto insurance contract. (We say "very nearly" because if the loss is smaller than the deductible, then the amount the insured owner pays does depend on the size of the loss.)

It is helpful to contrast the practice in automobile insurance with the practice in health insurance and health-care plans, where it is common to require copayments from the consumer for any services used. A consumer's choices about when to visit the doctor, whether to seek urgent care or to wait for a regular appointment, and so on, are all choices that affect the total level of cost incurred. The total level of cost incurred therefore provides information about how effectively the agent—in this case the consumer—has conserved scarce health-provision resources. As the theory predicts, the payments made by a health-insurance consumer therefore varies directly with the cost incurred by the health care provider.

The Incentive-Intensity Principle

The next step in the general analysis of incentive contracts is to determine how intense the incentives should be. In this step, we fix the information weighting parameter γ at whatever level the contract specifies (whether optimal or not) and let $V = Var(x + \gamma y)$.

> **The Incentive Intensity Principle.** The optimal intensity of incentives depends on four factors: the incremental profits created by additional effort, the precision with which the desired activities are assessed, the agent's risk tolerance, and the agent's responsiveness to incentives. The formula for the optimal intensity is: $\beta = P'(e)/[1 + rVC''(e)]$.

According to the incentive intensity principle, there are four factors that interact to determine the appropriate intensity of incentives. The first is the profitability of incremental effort. There is no point incurring the costs of eliciting extra effort unless