# 6

# MORAL HAZARD AND PERFORMANCE INCENTIVES

*Incentives are the essence of economics.*

*Edward P. Lazear*[1]

*The problem with corruption is that it tends to become the Problem of Corruption. Moral issues usually obscure practical issues, even when the moral question is a relatively small one and the practical matter is very great.*

*James Q. Wilson*[2]

Suppose that you are traveling along a highway when a dashboard light comes on indicating that your car is overheating. There is a service station nearby, so you drive your car there. The traditional analysis of this market situation is simple: There is some price to be paid to repair the problem; either you pay it and the car is fixed, or you decide to take your chances and decline to get the car serviced.

That account *might* reflect what would happen, but there are other possibilities. After beginning to work on the car, the mechanic might say, "Your radiator is shot. A new one will cost $500." If you are like most drivers, you have no idea whether the mechanic is being truthful or not. After all, it may be in the mechanic's interest to sell you a radiator, especially at that price. You face the same problem that decision makers in organizations frequently face: *When those with critical information have interests different from those of the decision maker, they may fail to report completely and accurately the information needed to make good decisions.* If the mechanic is lying to you, then both your interests and society's interest in efficiency are harmed. Your interests are harmed because the mechanic is profiting at your expense, and society's interests are harmed because productive resources have been wasted: A radiator that could have been repaired has instead been discarded.

Suppose that you agree to buy the radiator, wait an hour while it is installed, and then proceed down the highway another 100 miles toward your destination. You notice the overheating indicator on your dashboard lighting up again. Pulling into another service station, you learn that the new radiator was not installed correctly and it will cost you another $35 and another hour of waiting to have the job done right. *When buyers cannot easily monitor the quality of the goods or services that they receive, there is a tendency for some suppliers to substitute poor quality goods or to exercise too little effort, care, or diligence in providing the services.* Once again, both you and society are harmed. You, because you paid and waited twice for the same service, and society, because resources were wasted: It took two hours of mechanic time (and waiting time) to do a job that could have been done in one.

## THE CONCEPT OF MORAL HAZARD

These problems with the sale and installation of your radiator are examples of **moral hazard**, which is the form of postcontractual opportunism that arises because actions that have efficiency consequences are not freely observable and so the person taking them may choose to pursue his or her private interests at others' expense. The possibility of this sort of misbehavior is ruled out in the neoclassical model of markets considered in Chapter 3, where it was (somewhat implicitly) assumed that the transactions that people undertake are simple exchanges of goods and services with specific, well-understood, observable attributes, and that parties to a transaction can costlessly verify whether the terms of the transaction are being met.

### Insurance and Misbehavior

The term *moral hazard* originated in the insurance industry, where it referred to the tendency of people with insurance to change their behavior in a way that leads to larger claims against the insurance company. For example, being insured may make people lax about taking precautions to avoid or minimize losses. If the necessary precautions were known in advance and could be accurately measured and recorded, then an insurance contract could specify which precautions must be taken. However, frequently it is not possible to observe and verify the relevant behavior and thus it is not possible to write *enforceable* contracts that specify the behavior to be adopted. (The contract could call for the desired behavior, but how could the insurance company tell if the contract terms had been met?)

The kinds of moral hazard associated with insurance arise frequently in daily life. For example, you are likely to be much more careful in driving a rented car if you are financially responsible for all damage to the car than if you have purchased the Collision Damage Waiver and so are insured against the costs of dents and scrapes. Similarly, if you are covered under health insurance or belong to a Health Maintenance Organization (HMO), so that you are insured against all or most of the costs of visits to the doctor, you are likely to make greater use of medical services of all kinds: doctor

---

[1] "Incentive Contracts," in *The New Palgrave: A Dictionary of Economics*; Vol. 2 J. Eatwell, M. Milgate and P. Newman, eds. (London: The Macmillan Press, 1987), 744–48.

[2] "Corruption Is Not Always Scandalous," in *Theft of the City: Readings on Corruption in America*, J. A. Gardiner and D. J. Olsen, eds. (Bloomington: Indiana University Press, 1968), p. 29. Copyright © 1968 by the New York Times Company. Reprinted by permission.

visits, emergency room visits, prescription drugs, prenatal care, and so on. In each case, the fact that you are insured alters your behavior in ways that are costly to the insurer.

## Efficiency Effects of Moral Hazard

Although the term *moral hazard* has negative connotations, not all of the changes in behavior occasioned by insurance are socially undesirable, because some social interests may not be represented in the bargain between the insurer and the insured. For example, increased prenatal visits may result in healthier mothers and babies, consistent with society's goals. Moreover, the insurance company may not suffer any losses from moral hazard if it sets the insurance premium high enough to cover the extra costs. Still, moral hazard does impair people's ability to make mutually beneficial agreements and does often interfere with efficiency.

Moral hazard in insurance presents an efficiency problem because the extra benefits enjoyed by the insured on account of his or her changed behavior will often not be worth the costs. This happens because the insured decision maker does not look at all the costs and benefits associated with his or her decisions. Moreover, the inherent nature of insurance makes this inevitable. In the car rental example, you bear the full costs of the care you exert, but if you are fully insured, then being careful brings you no extra benefits. In the health care example, you get the benefits of the extra treatment you seek when insured, but bear little or none of the costs. Thus, you will tend to go to the doctor for minor ailments whose treatment seems worth your time, even if the total costs may exceed the benefits you receive.

The incentives to alter behavior would still not be a problem if it were easy to determine when the behavior were appropriate and to prevent excessive use. This sort of monitoring is often impossible, however, or at least very costly. There is no cost-effective way for the car rental company to observe the care you take. In the medical context, you will often lack the expertise to judge whether a particular visit is necessary, and it will not generally be in the doctor's interest to report that you have made an unnecessary visit. For this reason, moral hazard is an information problem: The difficulty or cost of monitoring and enforcing appropriate behavior creates the moral hazard problem. These difficulties mean that contracting is incomplete because there is no point to writing a contract specifying particular behavior when the desired actions cannot be observed and consequently the contract cannot be effectively enforced.

In terms of standard economics, the insurance has lowered the cost to you of something you value (doctor visits, not bothering to exert great care), and you thus "buy" more. More particularly, you now pay less than the full costs of extra units and so you purchase an inefficiently large amount. Putting the issue in these terms suggests that the behavior is not especially wicked and that the negative implications in labeling it "moral hazard" may be somewhat misplaced. More significantly, it suggests that this sort of behavior—and the efficiency losses it induces—might be quite widespread. Such is indeed the case.

## The Incidence of Moral Hazard

Moral hazard problems may arise in any situation in which someone (who may be a supplier, a customer, an employee, or anyone else) is tempted to take an inefficient action or to provide distorted information (leading others to take inefficient actions) because the individual's interests are not aligned with the group interest and because the report cannot easily be checked or the action accurately monitored. These problems are pervasive both in markets and in other forms of organization. Some doctors in the United States, for example, in an attempt to protect themselves from malpractice suits, practice "conservative medicine," ordering tests and procedures that may not be in the patient's best interests and, in any case, are surely not worth the costs (which are borne by the patient or the insurer—not by the doctor making the decision). Some firms may find it most profitable to make shoddy or unsafe products when quality is not easily observed. Security brokers may "churn" their clients' portfolios, encouraging them to trade more frequently than they really should because each additional trade generates commissions for the brokers. Automobile dealers may fail to mention the poor resale values or the higher than average repair costs of the cars they sell. Rented apartments may be less well maintained than owner-occupied ones because the renters do not get the full benefits of their efforts at maintenance. All of these examples are drawn from ordinary market experience.

Within organizations, an office employee may spend time during the day studying for an accounting exam, thinking about a new business idea that he or she hopes to pursue, or chatting on the telephone with friends when there is work waiting to be done. Factory workers may call in sick during hunting or fishing season, and when on the job they may exert the least care and effort they can get away with. Managerial employees may exaggerate the difficulty of their assignments in order to

### *Hidden Actions or Hidden Information?*

Although moral hazard and adverse selection usually seem quite distinct in textbook discussions, in practice it may be quite difficult to determine which is at work.

A radio story in the summer of 1990 reported a study on the makes and models of cars that were observed going through intersections in the Washington, D.C. area without stopping at the stop signs. According to the story, Volvos were heavily overrepresented: The fraction of cars running stop signs that were Volvos was much greater than the fraction of Volvos in the total population of cars in the D.C. area. This is initially surprising because Volvo has built a reputation as an especially safe car that appeals to sensible, safety-conscious drivers. Volvos are largely bought by middle-class couples with children. How then is this observation explained?

One possibility is that people driving Volvos feel particularly safe in this sturdy, heavily built, crash-tested car. Thus they are willing to take risks that they would not take in another, less safe car. Driving a Volvo leads to a propensity to run stop signs. This is essentially a moral hazard explanation: The car is a form of insurance, and having the insurance alters behavior in ways that are privately rational but socially undesirable.

A second possibility is that the people who buy Volvos know that they are bad drivers who are apt, for example, to be paying more attention to their children in the back seat than to stop signs. The safety that a Volvo promises is especially attractive to people who have this private information about their driving, and so they disproportionately buy this safe car. A propensity for running stop signs leads to driving a Volvo. This is, of course, essentially a self-selection story: the Volvo buyers are privately informed about their driving habits and abilities. Unless this selection imposes costs on Volvo, however, it is not adverse selection

Both stories are at least plausible. How would you go about testing which, if either, is correct? What other explanations seem plausible?

make their performance appear more impressive, or they may denigrate others' performance in order to improve their own chances of getting a particularly desirable assignment or promotion. Division executives may adopt policies that lead to high current performance that will be rewarded by bonuses and promotions, even though these policies will ultimately destroy the long-term profitability of the divisions they will have left behind. Senior executives may pursue their own goals of status, high salaries, expensive "perks," and job security rather than the stockholders' interests, and so they may push sales growth over profits, treat themselves to huge staffs and corporate jets, and oppose takeovers that would lead to their dismissal but would increase the value of the firm.

These examples do not involve insurance explicitly, but they have the crucial feature of insurance: The decision makers do not bear the full impact of their decisions. The doctor who orders extra tests benefits in terms of the reduced likelihood of successful malpractice suits but does not pay for the costs of the tests or suffer the discomfort they cause. The employees get paid whether they work hard or not, or at least they do not suffer a decrease in pay equal to the full lost value of what they could have produced. The difficulties of monitoring are what prevents them from bearing the full costs and benefits: The stockbroker's client does not have the expertise to tell if a trade is good for him or her or just for the broker; the employee's supervisor cannot freely determine whether he or she is thinking about company business or personal matters; and the stockholders cannot easily evaluate whether a particular executive action is in their interests or not.

THE PRINCIPAL-AGENT RELATIONSHIP Each of these examples can be cast in terms of an *agency* relationship. This term has come to be used in economics to refer to situations in which one individual (the **agent**) acts on behalf of another (the **principal**) and is supposed to advance the principal's goals. The moral hazard problem arises when agent and principal have differing individual objectives and the principal cannot easily determine whether the agent's reports and actions are being taken in pursuit of the principal's goals or are self-interested misbehavior. Agency relations in this sense are pervasive: The doctor is the agent of the patient, the worker is the agent of the firm, the CEO is the agent of the owners, and so on. As we will see later, however, moral hazard problems also arise in relationships where neither party can be considered the agent of the other but rather each is on an equal footing (as in a partnership).

## CASE STUDY: THE U.S. SAVINGS AND LOAN CRISIS

The key factors giving rise to moral hazard problems—divergent interests, decision makers being insured against some of the consequences of their actions, and monitoring and enforcement being imperfect—all feature centrally in one of the most spectacular moral hazard problems of recent times, the United States "savings and loan crisis."

### The Savings and Loan Industry

Savings and loan associations (S&Ls) are for-profit financial institutions that borrow money from the public in the form of deposits and then invest it by lending it out again, much like banks. The deposits of individual depositors in an S&L are insured by a U.S. federal government agency—until 1990, the Federal Savings and Loan Insurance Corporation (FSLIC). If, for some reason, the S&L could not repay the deposits, the FSLIC would. Government-provided insurance for bank deposits was instituted in the United States in the 1930s to protect depositors against bank failures. This insurance was also intended to reduce the likelihood of bank failures by eliminating "bank runs," which arise when depositors become fearful that their deposits may not be repaid, rush to withdraw their funds, and thereby bring on the failure they feared. The S&Ls, as well as the not-for-profit credit unions, were also provided with deposit insurance. The funds to pay for claims against the FSLIC came from charges levied on the insured S&Ls. The size of these premia were not linked to the riskiness of the S&Ls' portfolio of loans and other investments.

THE CRISIS IN THE 1980s Traditionally, the S&Ls were strictly limited in how they could invest their funds, with their primary investments being residential mortgage loans to local individuals secured by the homes they owned. During the 1980s many S&Ls turned to riskier investments, including loans on commercial real estate and high-yielding but very risky corporate borrowing called "junk bonds." As the commercial real estate market collapsed in several parts of the country, borrowers ceased payments on many of their loans, and the S&Ls were left holding property they could not rent or sell. Later, defaults by some corporations on their junk bonds undercut the value of all high-risk debt, further reducing the S&Ls' assets. As well, a plague of fraud spread through the industry. This proved a devastating combination: Over 500 savings and loans slipped into bankruptcy. The FSLIC's reserves were inadequate to cover its promises to protect depositors, and U.S. taxpayers are now having to foot the bill, which is measured in the hundreds of *billions* of dollars! What led the S&Ls to make such risky investments? What led to the increase in fraud? Could it all have been prevented? Who is to blame?

THE CAUSE: MORAL HAZARD The very design of the deposit insurance program, together with lax regulation, led to a costly problem of moral hazard in the management of the savings and loans. In brief, deposit insurance and low capital requirements (the amount of the S&L owners' own money at risk) encouraged excessive risk taking by relieving the S&Ls of the responsibility for poorly performing investments while allowing them to gain when the investments prospered. The insurance also relieved the depositors of the usual responsibility of investors to keep tabs on those who hold their money. This encouraged both risk taking and fraud. Insurance could be economically provided only so long as other regulatory policies were able to prevent the S&L managers from making reckless investments and engaging in self-dealing. In the early 1980s, however, the regulations controlling the sort of investments the S&Ls could make were relaxed. At the same time, the amount of insurance afforded to each depositor was increased and the resources devoted to enforcing the relaxed regulations were reduced. The whole system inevitably broke down.

### Deposit Insurance and Risk Taking: An Example

The S&Ls made risky investments in part because the government insurance scheme made those investments profitable for the owners of the S&Ls. To see how insurance creates these incentives, let us study an example. To make the calculations simple, we set the interest rate paid to depositors at zero. The principles we deduce from this example apply to any other interest rate as well.

Suppose the owner of the S&L has full authority over how to invest the deposits. The owner can choose between two possible investments, labeled "safe" and "risky." Actually, both investments in our example have some uncertainty about their returns but, as Table 6.1 shows, the safe investment has less variation in its returns and can never actually lose money.

The safe investment requires an initial outlay of $100 and returns either $100 or $110, each with a 50 percent probability. We call it "safe" because it always returns at least the initial outlay of $100. On average, it does even better, returning the initial outlay plus $5 more.

The risky investment is a lemon. There is a 50 percent chance that it will return only $65, far less than the initial outlay of $100. On average, it loses $5. This is a

Table 6.1 **Description of Investment Opportunities**

| | Safe | Risky |
|---|---|---|
| Initial outlay | 100 | 100 |
| High return | 110 | 125 |
| . . . probability | .50 | .50 |
| Low return | 100 | 65 |
| . . . probability | .50 | .50 |
| Expected return (gross) | 105 | 95 |
| Expected return (net) | 5 | −5 |

bad investment that, in a well-functioning system, would not be undertaken because it wastes social resources. As we shall see, however, the way the S&L is financed and insured can create differences of interests between the various parties whose resources are at stake, and lead the managers to undertake the risky investment.

The funds an S&L has for lending and investing come from two main sources: deposits and the capital supplied by owners. United States federal regulations in the 1980s required the owners of an S&L to provide capital equal to about 3 percent of the total value it invests. For our example, we suppose that of the initial outlay of $100 required for the investment, $97 comes from the depositors (insured by the FSLIC) and $3 comes from the owners of the S&L. The depositors have first priority on any proceeds from the investment. This means that if the proceeds from the investment are more than $97, the depositors must be paid in full. If they are less than $97, the depositors get whatever money is available, the owners get nothing, and the FSLIC pays the difference between the available funds and depositors' $97 claim.

SHARING THE RISKS Now let us examine the distribution of the costs and benefits if the safe investment is made. The term *gross return* refers to all the income received from the investment, without regard to the initial outlay. In Table 6.2, the gross return earned by the owners is $13 if the investment works out well and $3 if it works out badly. *Net return* refers to the gross return minus the initial outlay. Because the owners have an initial outlay of $3, their net returns are either $10 or $0. In expectation, they get $5. In any event, the depositors just get their money back, netting zero (the interest they were promised), and the FSLIC neither receives nor disburses any funds.

When the safe investment is made, the FSLIC is never needed to "bail out" the S&L, that is, to help it meet its obligations. The situation is quite different when the risky investment is made, however. Table 6.3 shows how the returns on the risky investment might appear to the various parties.

Table 6.2 **Analysis of the Safe Investment**

| | Depositors | Owners | FSLIC | Total |
|---|---|---|---|---|
| Initial outlay | 97 | 3 | 0 | 100 |
| High return (gross) | 97 | 13 | 0 | 110 |
| Low return (gross) | 97 | 3 | 0 | 100 |
| Expected return (gross) | 97 | 8 | 0 | 105 |
| Expected return (net) | 0 | 5 | 0 | 5 |

Table 6.3 **Analysis of the Risky Investment**

| | Depositors | Owners | FSLIC | Total |
|---|---|---|---|---|
| Initial outlay | 97 | 3 | 0 | 100 |
| High return (gross) | 97 | 28 | 0 | 125 |
| Low return (gross) | 97 | 0 | −32 | 65 |
| Expected return (gross) | 97 | 14 | −16 | 95 |
| Expected return (net) | 0 | 11 | −16 | −5 |

The final column of Table 6.3 repeats the description of the risky investment already given in Table 6.1. The first three columns show how the initial outlay and returns are divided among the three parties. Notice that when returns are high, the owners of the S&L enjoy exceptionally high profits. In contrast, when returns are low and there is not enough money to pay the depositors' claims, the FSLIC bears exceptional costs to pay off the depositors. All the owners of the S&L lose is their $3, with the FSLIC absorbing the rest of the loss. The S&L owners benefit from risk taking, whereas the FSLIC suffers from it.

THE WINNERS AND LOSERS The bottom line of Table 6.3 further clarifies the matter. Although this investment is a lemon overall, with an expected net return of −$5, it generates an expected net return of $11 to the owners. This is far more than the $5 the owners could expect to get from the socially preferable, safe investment. The FSLIC, which bears the losses when the investment returns are too low to cover deposits, now suffers an expected loss of $16, the difference between the owners' expected returns and the total returns on the investment. The depositors, who are insured, always get all of their money back.

According to the bottom line of Table 6.3, the expected net returns of all the parties add up to the total expected net returns of the investment. Because the depositors always get zero, each dollar of expected loss imposed on the FSLIC shows up as another dollar of expected profit for the owners! In choosing among investments with equal expected returns, the owners will prefer the riskier ones because they expect to profit most when the FSLIC's expected losses are largest.

## Incentives for Risk Taking with Borrowed Funds

The financial motivation for the S&L owners to make risky investments is now clear: The riskier the investment, the higher the expected losses for the FSLIC and the greater the expected profits for the S&L. Still, our story of how the government botched its regulation of the S&Ls is incomplete. The problem is that the owners would have the same motivation to make risky investments even if the FSLIC were taken out of the picture! With the FSLIC gone, the losses would fall on the depositors instead of the government agency. Still, the owners would be the ones to benefit from the risky investment if things go well, and someone else would be left to bear the losses. Thus it might appear that if the government eliminated deposit insurance, the only thing to change would be that risks are shifted from the government agency to the depositors. *So long as investments are financed by borrowing, the borrowers will always have an incentive to undertake riskier investments than the lenders would want.* Of course, the whole point of a savings and loan institution is that depositors leave their money there for the S&L to invest; an S&L always is a borrower from its depositors, so the problem always exists. It may seem, therefore, that we have unfairly identified the FSLIC as being the root of the problem.

### *The Case of Seapointe Savings and Loan*

Seapointe Savings and Loan was founded in 1985 in Carlsbad, California, a suburb of San Diego. Like other S&Ls, Seapointe's primary business was to make home mortgage loans. According to its business plan: "At no time will management presume to outguess the marketplace nor risk the net worth of the institution in an attempt to 'make a killing' for the sake of short-term earnings."

When Seapointe failed in 1986, however, it had never even hired a loan officer. Instead, it had sold "naked call options." That is, it had sold a promise to deliver $10 billion of bonds that it did not own, at the buyer's option, at a given future date. Selling call options is a common practice, but selling them without actually owning the bonds makes the options "naked"; it exposed Seapointe to the risk of an actual cash loss if bond prices were to rise. And rise they did, so that Seapointe was forced to pay the difference between what the bonds cost at the delivery date and the price it had promised. Seapointe lost $24 million on this one transaction—75 percent of its assets—leaving the FSLIC responsible to repay the institution's depositors. If Seapointe had won its bet and bond prices had fallen, the buyer would not have exercised its option to demand delivery of the bonds at the promised price, and Seapointe would have kept the amount it was paid for its promise. Seapointe would have 'made a killing' for its owners.

Source: Charlotte-Ann Lucas, "How an S&L Gambled Off Its Deposits Within a Year," *San Francisco Examiner* (December 2, 1990), A-1.

MONITORING BORROWERS To gain a deeper understanding of the issues, we must look beyond the S&L industry to find related business problems. The problem of depositors who are, in effect, lenders to the S&L is by no means unique in business. Lenders exist throughout the business world. However, they take precautions to ensure that their money is not squandered, stolen, or put at unnecessary risk by those who have borrowed it. A bank that lends you money will ask you about your financial condition and about what you intend to do with the loan proceeds. It will run a credit check, demand collateral, and often require regular payments of the interest and part of the principal. If yours is a home loan, it will demand a legal interest in the property, and if it is a car loan, the bank will keep title to the car until the loan is repaid. Those who lend to firms frequently impose similar conditions. They examine the firm's financial condition and credit history, put restrictions on how their funds may be used, and often require business plans, collateral, and periodic financial statements. Correspondingly, careful scrutiny by depositors is the mechanism by which an unregulated and uninsured S&L might be kept from making irresponsible investments or defrauding its investors.

Why, then, did S&L depositors not take the same precautions as other lenders? Because doing so was costly and, anyway, the deposits were insured! The insurance itself made the depositors willing to supply huge sums to S&Ls without the usual checking of creditworthiness or monitoring of performance that accompanies other large loans. For the FSLIC to protect itself against huge losses, it thus needed to regulate the S&Ls, monitoring their activities, restricting their investments, and ensuring that they maintained adequate capital both to guard against unlucky investment outcomes and to ensure that the owners would suffer a significant loss if the organization should fail. The government failed to do this through the 1980s and suffered from the resulting moral hazard problem.

## The Perverse Effect of Competition

The moral hazard problem in the S&L industry was actually intensified by the effects of competition. Normally we think of competition, which tends to drive out those executives who are unwilling to take the profit-maximizing actions, as promoting efficiency. In the context of the S&L industry in the 1980s, however, competition had a perverse effect. Many conservative S&L executives had no choice but to gamble on risky investments if they were to survive in the circumstances we have described.

Think about how the system works. The problems may have all begun with a few S&Ls that directly saw the chance to exploit the deposit insurance system by moving into more risky investments. To do so, they needed to expand their deposit base. The only quick and sure way to attract substantial new deposits is to offer higher interest rates to depositors. Thus, S&Ls seeking an influx of money to expand investments offered higher interest rates than did their competitors. The government's increasing the amount that was insured (from $40,000 per account to $100,000) made these higher rates very effective, as large investors—including many firms—deposited their money in the aggressive S&Ls.

Now, the other S&Ls began to feel the heat. Deposits were being drawn out of their doors and given to their competitors. To stay in business, some of these others also raised the rates paid on their deposits. For some, given their operating costs, these rates were higher than they were able to pay using just their normal, relatively safe investments in residential real estate. Therefore, they too were driven to riskier investments which, if they worked out well, would enable the company to pay the promised interest rates and still make a profit.

Despite the spiraling competitive pressure, some S&Ls may have held out, making only safe investments. They either offered lower interest rates to depositors and so faced a crisis of falling deposits, or they matched the competitive, higher interest rates and suffered losses as the income from their loans fell short of what was needed to pay their costs and other obligations. This became especially significant in 1979 and 1980 when a change in monetary policy by the Federal Reserve (the U.S. central bank) caused interest rates to shoot up throughout the economy. Many S&Ls had much of their money tied up in long-term, fixed-rate mortgage loans, the rates on which suddenly were less than what they were having to pay for their deposits. As well, the collapse of the real estate market in Texas when oil prices fell in the mid-1980s had a special adverse effect on the S&Ls in that state.

Many of the endangered S&Ls became prey for aggressive entrepreneurs, who bought the failing companies for low prices and tried to make them profitable by radical means—offering very high rates on $100,000 "jumbo" deposits and investing in risky commercial real estate, "junk bonds," and other similar ventures. Investors continued to place their deposits with these financially troubled companies because the deposits were federally insured.

It is easy to see how competitive pressure forced out many of the more conservative S&L executives throughout the industry. Those who were unwilling to make the risky investments were often driven out of business. The big loser was the FSLIC—and the taxpayer.

## Fraud in the S&Ls

The story of the savings and loan industry as told here leaves out many important details. Risky investments that went bad did not alone deplete the capital of so many S&Ls. Outright fraud was also responsible. News accounts report that the top officers at savings and loan institutions made loans to themselves, their other companies, their friends, and their family members at reduced interest rates and without adequate collateral. They paid large dividends to investors and generous salaries to themselves and their relatives, even as their firms were sliding into bankruptcy. They concealed bad loans by lending more money to the borrowers so that they could afford to make payments on older loans. These activities are tantamount to stealing funds from the S&L, or, because of the insurance, from the taxpayers. Government investigators have found that there was fraud in at least 25 percent of the cases of S&L bankruptcy.

An analysis of the problem of fraud in the savings and loan industry would be quite similar to our analysis of the adoption of risky investments. Throughout the economy, people entrust their funds to the management of others. They protect against fraud and against excessive risk taking in the same sorts of ways: by monitoring performance, hiring auditors, writing restrictive rules into the organization's charter about what activities are allowed, and so on. They retain enough control of the management's activities to dismiss errant executives. The depositors at a federally insured savings and loan did not engage in these costly activities because the deposits were backed by an agency of the U.S. government. No private investor has an incentive to protect the federal insurance agency against fraud; the government regulators have to do that for themselves.

## Who's To Blame?

This example of the S&L crisis is remarkably rich, for it involves moral hazard on the part of three distinct groups. First, and most obvious, are the S&L owners who took excessively risky investments or committed fraud. Second are the depositors who failed to monitor the S&Ls because their deposits were insured. The third group consists of the politicians—in both the legislative and the executive branches—who favored the industry at the expense of the general taxpayer. These politicians raised the amount of insurance provided by the FSLIC, thereby making it easier to attract large deposits. They relaxed the regulations on the S&Ls and did not provide for an offsetting increase in monitoring. Furthermore, when the S&Ls were first headed for financial trouble, politicians blocked the regulators from intervening to protect the FSLIC and the taxpayers. Possibly some of these politicians were motivated by a genuine belief that the actions being taken were truly in the general interest. However, the huge campaign donations made by some S&Ls to various of these politicians certainly raise the possibility that the politicians were pursuing their own interests and expected to get away with it because the public's monitoring of them is so imperfect.

### PUBLIC VERSUS PRIVATE INSURANCE

The savings and loan crisis is fundamentally the result of moral hazard that arises from the existence of the deposit insurance provided by the FSLIC. Yet this insurance provides valuable social benefits as well. Small savers need not lose sleep worrying about the safety of their deposits, nor do they have to incur the very real costs of monitoring the institutions to which they have entrusted their life savings. The desirability of deposit insurance, combined with the obvious problems that the government-provided program experienced, has led some commentators to suggest that such insurance should be privately provided. There are clear difficulties in designing and implementing such a program, and the failure of a number of small private deposit insurance funds in the state of Rhode Island in 1991 may have removed some of the allure of this idea. Still, it is clear that government insurance programs do seem to have inordinate difficulties with moral hazard.

## Other U.S. Government Insurance and Guarantee Programs

The crisis in the savings and loan industry is symptomatic of problems with many other government insurance and guarantee programs in the United States. Together these programs are estimated to involve insurance and loan-guarantee commitments of more than $5 *trillion*, which is almost twice the U.S. national debt and five times the level of yearly federal government spending.[3] Here are just a few examples.

THE PENSION BENEFIT GUARANTY CORPORATION The Pension Benefit Guaranty Corporation (PBGC) was established by the Employee Retirement Income Security Act of 1974 (ERISA) with the intent, according to its advocates, of ensuring that promises of retirement benefits made to working people by their employers would be honored. The PBGC is obliged to take over plans that are terminated without sufficient funds to pay the promised benefits. It collects what it can from the company that terminated the plan and uses the proceeds to pay the pensioners. To finance the remainder, it collects a tax called an *insurance premium* that is imposed on other pension plans.

To avoid transferring huge pension liabilities to the new government agency, ERISA provided certain minimum funding standards for pensions and held employers liable in part for their pension promises. The act allowed dramatic underfunding of certain kinds of plans, however, particularly multiemployer plans run by trade unions for their own members, but also some corporate pension plans. Predictably, many of these plans aggressively expanded benefits beyond what the limited funds available could possibly justify. Later, some of these plans shut down, saddling the PBGC with the liability to pay the promised benefits.

THE FEDERAL CROP INSURANCE CORPORATION The Federal Crop Insurance Corporation (FCIC) was established in 1939 to protect farmers against crop losses caused by the vagaries of weather. Yet the FCIC has relatively few inspectors and has been raked by fraudulent claims. Some farmers have defrauded the FCIC by claiming crop losses in one name and selling the harvest in another. Some crop insurance policies have been sold after the claimed loss occurred. In one case a claim was filed for crops that were planted 30 days after the freeze that had triggered the insurance payment.[4]

Even when fraud is not a factor, moral hazard arises. Insured farmers are tempted to take greater risks by planting less hardy or more water-hungry crops than would otherwise be prudent. If the weather turns out to be warm or rainfall turns out to be plentiful, the farmers profit; if not, the government insurance program pays.

THE GOVERNMENT NATIONAL MORTGAGE ASSOCIATION The Government National Mortgage Association (GNMA) exists to make it easier for homeowners to obtain mortgage loans. It does this in several ways, most prominently by providing insurance to lenders against defaults on the mortgages they write. Along with the development of mortgage-backed securities (see Chapter 1) came mortgage brokers who received commissions of $30 to $45 per $1,000 of the loans they wrote, while maintaining capital of as little as $0.30 per $1,000 of outstanding mortgages.

It is profitable for these brokers to write very risky mortgages because they receive a large commission but have little capital to lose in the event of a default; GNMA

[3] "Government Waste: Where's Nanny?" *The Economist* (January 6, 1990), 31.

[4] Bruce Ingersoll, "Crop-Insurance Fraud and Bungling Cost U.S. Taxpayers Billions," *The Wall Street Journal* (May 15, 1989), A-1.

picks up the tab. To protect itself, GNMA specifies limits on the ratio of the loan amount to the appraised value of the property, so that it will have adequate collateral. Appraisals are subjective, however, and unscrupulous brokers have on occasion vastly exaggerated the value of properties in order to justify large loans, leading to default and large losses paid for by the taxpayer while the borrower and broker profit.

STUDENT LOANS Student loans provide another example of how guarantees affect costs. Federal government guarantees enable many students to obtain loans on more favorable terms than would otherwise be available. Normally, when banks make loans, they take measures to ensure that the loans are collectible and they are aggressive in collecting from debtors. The incentives to ensure collectibility are blunted by government guarantees. Unsurprisingly, a huge proportion of guaranteed student loans are never repaid. Just tracking who the debtors are often exceeds the government's limited capacity to administer these loans.

### Private or Public Insurance?

Although moral hazard problems are present in both government and private sector insurance programs, the problems do seem to be less severe in the private sector. In part, this is because private corporations cannot sustain such huge losses for long without going bankrupt, and they often cannot rely on the taxpayers to pay for their financial ineptitude. But the more limited difficulties with moral hazard in private insurance is not due entirely to any special merit of private insurance programs; it is partly a result of the private sector's unwillingness to undertake socially desirable insurance programs in which the costs associated with moral hazard are high.

The money-making objective of private companies is quite different from the objectives of a government agency. No bureaucrat is well positioned to eliminate an unprofitable deposit insurance program that protects the life savings of small depositors. Even a legislature would have difficulty making such a decision. Nor is it clear that elimination of such a program is socially desirable. In contrast, most private firms would have little trouble deciding to terminate the program if the losses being suffered were large.

On average, private-sector insurance programs do seem to be better managed than are their government counterparts. Some of the losses suffered by the various government programs could have been avoided by having more inspectors or tighter regulations. The "output" of inspectors, however, is not easily measured and so, especially during periods of large budget deficits, there is an attractive short-term economy to be achieved by reducing the number of inspectors on the government payroll. Short-sightedness and poor management of this sort certainly seem to have been a factor in the savings and loan crisis.

Yet moving insurance to the private sector is no cure. Moral hazard can manifest itself there in some remarkable ways as well.

### Moral Hazard in Private Life Insurance

In a life insurance contract, the insured's designated beneficiary is paid an agreed-upon sum of money when the insured person dies. All life insurance contracts issued in the United States have certain standard provisions, one of which deals with death by suicide. Essentially, life insurance contracts provide that the insurance company will pay off on the policy to the beneficiary after the insured person dies by suicide only if the suicide occurs after a certain period of time has elapsed from the time the policy was issued. In the United States, this exclusion period is always either 12 or 24 months. Life insurance statistics show that the suicide rate is lowest in the twelfth and twenty-fourth months after a policy has been issued and highest in the thirteenth and twenty-fifth. The inference seems unmistakable: People postpone their suicides to allow their beneficiaries to collect the life insurance proceeds.

## MORAL HAZARD IN ORGANIZATIONS

Moral hazard was first identified in the insurance context, and some of its most spectacular manifestations are still found there. For understanding organizations, however, it is important to recognize the point made earlier—that moral hazard is a very common phenomenon that affects a wide array of transactions and that attempts to deal with moral hazard account for many of the particular institutional arrangements we see, both in markets and within organizations. Indeed, the very boundary between these two forms of organization is often a response to moral hazard concerns.

### Moral Hazard and Employee Shirking

An important instance of moral hazard arises in employment relationships, where employees may shirk their responsibilities. Frederick Taylor, the "father of scientific management," once wrote: "Hardly a competent worker can be found who does not devote a considerable amount of time to studying just how slowly he can work and still convince his employer that he is going at a good pace."[5]

Evidence of the importance of moral hazard in the employment relationship is the frequency with which firms give employees incentive or performance contracts. These arrangements tie the employee's compensation to various measures of performance, and are meant to motivate effort, creativity, care, diligence, loyalty, and so on. Examples include pay tied to output, such as piece rates for manufacturing workers or bonus clauses that reward unusually large numbers of touchdown passes caught by football players; pay linked to sales, such as salespeople's commissions; pay linked to productivity improvements, as under "Scanlon Plans"; and various ways of linking pay and profits, including employee stock ownership plans, the Japanese practice of paying workers an annual bonus tied to firm profitability, and many executive compensation schemes. When well designed and well administered, these sorts of arrangements can be effective in promoting the desired behavior. Although clear communication to employees of what it is that the employer values is partly responsible for this effect, direct financial incentives are the key.

To see that these arrangements are evidence of moral hazard, note that the firm is not paying directly for what the employees are supplying but instead uses a *proxy* for it. What is actually being supplied are such things as the employees' intellectual and physical effort. What is paid for are the *results* of these inputs—sales and touchdown passes, for example. The amount and quality of the employees' efforts are difficult to monitor directly, whereas the results of their efforts may be more easily observed. Thus, rather than trying to pay for unobservable effort directly, the firm attempts to motivate employees to choose to work harder or better by rewarding outcomes that are more likely when they behave in the desired way.

AIR TRAFFIC CONTROLLERS: AN EXAMPLE[6] Air traffic controllers are charged with maintaining air safety by keeping airplanes in flight at specified, safe distances from one another. They use radar to track flights and radio to direct the pilots.

Federal government employees in the United States in the 1970s, including air traffic controllers, were covered by a disability program, the Federal Employees'

---

[5] Frederick Taylor, *The Principles of Scientific Management* (New York and London: Harper, 1929).

[6] This section is based on Michael E. Staten and John Umbeck, "Information Costs and Incentives to Shirk: Disability Compensation of Air Traffic Controllers," *American Economic Review*, 72 (December 1982), 1023–37.

Compensation Act. If they were unable to work because of a disability that was the result of their jobs, they were entitled to a receive a fixed percentage of their pay for the duration of their disability. This tax-free payment could be as high as 75 percent of the base salary. Given the tax rates of the period, a disabled worker might actually receive a higher take-home income under the disability program than when working.

In order to collect on a disability claim, the injury had to be shown to be both disabling and work related. The injury did not need not to be physical; stress-related disorders that prevented the employee from working would qualify. To control unjustified claims, the injury report had to be supported by a statement from a physician certifying the disability, another from the employee's supervisor describing the events leading up to the injury, prior symptoms, and the work environment, and first-hand reports from coworkers.

Air traffic controllers, whose jobs were viewed as unusually responsible and stressful, would have had a relatively easy time making a claim for disability based on nervous or emotional disability. Moreover, certain changes in 1972 and 1974 in the rules governing disability claims made claiming disability even more attractive for controllers. The 1972 changes provided for retraining for second careers for those air traffic controllers who were found to be disabled, even if the disability were not job related. The 1974 rule changes made monitoring false claims generally more difficult and made catching a fake stress-related claim especially difficult. If moral hazard were a problem among controllers, then these changes should have led to increased incidence of fake disabilities ("punching out") and an increase in claims, especially in the number of psychologically based claims.

In fact, the number of disability claims *did* rise with the initiation of each program, more than doubling in 1974 and continued to rise. The largest percentage increase following the 1972 change was in psychological and psychiatric illness, such as stress-related disabilities; it was largest by a factor of three. (Unfortunately, the data did not permit identification of the mix of claims following the 1974 change.)

More striking, however, was the apparent impact of the 1974 change on job performance. A controller who wanted to fake a claim for stress-related disability needed to show the disability was job-related to collect, and the examiners were directed to look for specific events on the job that either could have contributed to the stress or that were symptoms of the disability. This created an incentive to manufacture on-the-job incidents that could have caused the stress and that might also indicate that the employee was no longer capable of doing the job. The natural candidate here was a "separation violation," in which planes for which the controller was responsible came too close to one another.

The Federal Aviation Authority keeps track of two sorts of separation violations: System Errors and Near Mid-Air Collisions. The former represent any violations of the standard separation requirements; the latter are much more serious and directly life threatening. Because either sort of violation would do equally well for the purposes of filing a claim, a controller who did not want to cause unnecessary danger would be much more willing to generate a minor violation than a near collision. In fact, the number of Systems Errors jumped significantly after the 1974 change, but there was only a small, statistically insignificant change in the number of Near Mid-Air Collisions. Furthermore, the increase in Systems Errors tended to occur not when traffic was particularly heavy, as you might otherwise expect, but when it was relatively light and the controller could cause the "needed" violation at minimal risk.

Finally, a controller considering punching out had to decide when in his or her career to do so. Various factors, especially eligibility for retraining and an effective dependence of the amount of disability pay on years of service, made it much more attractive to punch out after five years of service. Before the 1974 changes, controllers with less than five years' experience (who presumably were relatively inexperienced and more prone to mistakes) and, to a lesser extent, those with more than ten years' experience (who were more likely to suffer from actual "burn-out") were responsible for most of the System Errors made. By 1976 personnel in the five-to-ten-year range were committing over 50 percent of the errors, although this group accounted for less than 30 percent of the total number of controllers. All this, then, is striking evidence of moral hazard.

## Managerial Misbehavior

The senior executives of corporations are charged with advancing the interests of the stockholders, who are the owners of the company. They are supposed to be overseen in this duty by the board of directors, who are elected by the stockholders and who are empowered to represent them in voting on major corporate decisions and setting the executives' compensation. Thus, both the executives and the board members are considered the agents of the stockholders.

Over 50 years ago, Adophe Berle and Gardner Means argued that the dispersed holdings of stock across a multitude of small investors had created an effective **separation of ownership and control**, with no individual stockholder having any real incentive to monitor managers and ensure that the officers and board were running the firm in the owners' interests.[7] Although this claim long remained highly controversial, evidence that has accumulated indicates that managers often do fail to promote the interests of the stockholders effectively.

The problem typically is not that the executives are lazy and do not work hard enough. Corporate executives put in remarkably long hours of very intense effort. Rather, the complaint is that they pursue goals other than maximizing the long-run value of the firm. Critics claim that executives invest firms' earnings in low-value projects to expand their empires when the funds would be better distributed to the shareholders to invest for themselves. They are alleged to hang on to badly performing operations when other teams of managers could run them more profitably or even when the operations are irredeemable losers. With the connivance of their hand-picked boards, they pay themselves exorbitantly and lavish expensive perquisites upon themselves. They resist attempts to force more profitable operations, especially by resisting takeovers that threaten their jobs. All these alleged misdeeds serve the interests of the managers themselves (and perhaps the interests of other concerned constituencies, such as employees), but not the interests of the firms' owners.

HOSTILE TAKEOVERS AND MANAGERIAL MISBEHAVIOR During the 1980s a wave of **hostile takeovers** occurred in the United States and to a lesser extent in the United Kingdom and Canada. A hostile takeover is the acquisition of enough of the shares in a company to give a controlling ownership interest in the firm, where the offer to acquire the firm is opposed by the target company's executives and directors. Successful hostile takeover attempts generally resulted in replacement of the target firms' senior management and the naming of new boards of directors. The buyers in these transactions were called "corporate raiders" (as well as many less complimentary things) by the managers of the target firms who fought to maintain the companies' independence. In this context, "independence" means the firms' continuing under their current managers and boards with the existing ownership. This ownership was

[7] Adophe Berle and Gardner Means, *The Modern Corporation and Private Property* (New York: MacMillan, 1932).

typically quite diffuse—individual small shareholders, plus pension plans, insurance companies, and mutual funds.

Many observers have interpreted the hostile takeovers as a corrective response to managerial moral hazard: The takeovers, it is claimed, were intended to displace entrenched managers who were pursuing their own interests at the expense of the stockholders. Whether this is the case or not, the huge profits that were generated in these transactions raise questions about how effectively incumbent managers were maximizing the values of the companies they ran.

The prices paid for the stock of firms in hostile takeovers in this period on average represented a 50 percent premium over the target's original market value. For example, just before Mobil Oil launched its bid for Marathon Oil in 1981, Marathon's stock was trading at $63.75 a share. Mobil offered $85, and eventually raised its offer to $126 before Marathon was acquired by U.S. Steel (now called USX). Similar premia were paid in hundreds of cases. In aggregate from 1977 through 1986, shareholders selling their stocks in hostile takeovers realized an estimated gain of $346 billion (in 1986 dollars).[8] The takeover premia appear to be evidence of managerial incompetence or moral hazard to the extent that this original market value represented the discounted profit stream that savvy investors expected the firm to generate under its original management, whereas the takeover price reflected the firm's value under the new ownership.

The takeover premia are not conclusive evidence that the managers were poor stewards, however. Perhaps there was systematic overbidding by raiders who suffered delusions about their managerial abilities or were using other peoples' money to expand their own empires. Perhaps the new owners expected to reap gains at the expense of other stakeholders (especially current managers and workers), where these private gains are not increases in efficiency but the returns to violating explicit and implicit promises. Or perhaps the stock market was systematically underestimating the target firms' prospects before the takeover attempts were launched.

We examine these issues in more detail in Chapter 15, but one feature of the takeover wave does seem to be a clear manifestation of misbehavior: The adoption of a *poison pill* provision without an approving shareholder vote.

POISON PILLS Poison pills are takeover defenses.[9] They involve creation of special securities that give certain rights to their holders in the event that a raider acquires more than a specified fraction of the shares in the firm. Most commonly these rights are to buy shares in the target (or, if the takeover occurs, the acquiring) firm at very low prices. They work as takeover defenses because they vastly increase the cost of acquiring the firm. Although they are often labeled as "Shareholder Rights Plans" by managers trying to sell them to their shareholders (who receive some or all of the special securities), they in effect remove the ability of the owners of the firm to sell their shares to a buyer that their nominal agents—management and the board—do not like.

If the shareholders adopt such a scheme, it is largely their business, and it may well be in their best interest: Takeover defenses, if not strong enough to make takeovers impossible, may improve the stockholders' bargaining position and raise the price they ultimately receive if an acquisition does occur. Boards of directors can adopt poison pills without shareholder approval, however, and they often did so during the 1980s when they (or management) became nervous about possible takeovers. Doing so seems to be simply an expropriation of the shareholders' property by those who are supposed to be looking out for and serving their interests. Moreover, the empirical evidence is that adopting a poison pill typically reduces the firm's share value, and the firms that have adopted poison pills tend to be ones where managers and board members hold very few of the company's shares.[10] This evidence further supports the view that the adoption does not serve shareholder interests.

## Moral Hazard in Financial Contracts

A common sort of moral hazard problem arises when different individuals have differing claims on the financial returns from an investment. We have already seen an instance of this in the savings and loan crisis, but other examples abound and account for many elements of the form of financial contracts.

DEBT, EQUITY, AND BANKRUPTCY Many firms are financed by a combination of **debt** and **equity**. The debt holders—banks, the purchasers of the firm's bonds, input suppliers who offer credit—are lenders. They provide cash in return for a promise to be repaid a fixed amount (perhaps with interest) at a later date. The equity holders get to keep whatever profits are left after paying the debt obligations. In a corporation, the equity is lodged with the stockholders, who elect the board of directors to represent their interests in setting policy and in hiring managers to run the firm. In a partnership or sole proprietorship, the partners or owners are the equity claimants. Absent serious managerial moral hazard, we should expect that the firm will be run in the interests of the equity holders. This is not necessarily consistent with the interests of the firm's creditors.

We already saw one form of a conflict of interest between equity and debt in the savings and loan example. Equity holders will favor riskier investments than the firm's creditors would want. The reasoning is exactly the same as was developed there: The equity holders win big if the investments work out, whereas the debt holders just get their promised fixed payment, and if the investment loses money, some of the loss may fall on the creditors who are not fully repaid.

As we also noted in the savings and loan example, lenders take measures to protect themselves against the potential moral hazard problem that arises if the firm is run to maximize the value of its stock. They do credit checks, they demand collateral, they monitor performance (in part by requiring ongoing repayment), and they may structure the loans so that they can demand immediate repayment if they get nervous about the firm's ability to pay. Moreover, in some countries, the firm's bankers are normally named to the board of directors of the corporation, where they can more easily monitor their investments. As well, the bond holders may insist on covenants in the debt contract that limit the sort of actions the firm can take and the amount of additional borrowing it can undertake.

Despite all this, sometimes the loan cannot be repaid, or at least a scheduled payment is missed. In these circumstances, the lenders can force the firm into bankruptcy. Bankruptcy can be seen as an institutional arrangement to protect the value of assets. Once a firm is forced into involuntary bankruptcy, the creditors gain many of the decision rights that normally belong to equity. This prevents more of the resources available for meeting the debt obligations from being squandered. It thus

---

[8] See Michael Jensen, "Takeovers: Their Causes and Consequences," *Journal of Economic Perspectives*, 2 (1988), 21–48.

[9] For a discussion of takeover defenses, see J. Fred Weston, Kwang S. Chung, and Susan E. Hoag, *Mergers, Restructuring, and Corporate Control* (Englewood Cliffs, NJ: Prentice Hall, Inc., 1990), Chapter 20.

[10] Michael Jensen, "Takeovers: Their Causes and Consequences," 21–48.

makes people more willing to lend money than they otherwise would be and encourages the efficient allocation of financial capital. Moreover, to the extent that managers lose in a bankruptcy—because their jobs, their perks, and their pensions may disappear—the threat of bankruptcy may serve as a check on managerial moral hazard vis-à-vis stockholders' interests.

Under U.S. tax laws, the payments that a corporation makes as interest on its debt are tax deductible, whereas dividend payments on its stock are not. Because both are payments by the corporation for the use of capital, this might suggest that the firm would gain by financing itself overwhelmingly with debt. The attendant moral hazard problem is one reason why this would not be automatically attractive, however. As the fraction of the firm financed by debt increases, there is a growing incentive for equity holders and the managers who represent them to take risks. This means that at very high levels of debt to equity, the firm will have to pay very high interest rates, put up extremely large amounts of collateral, and accord lenders extensive control rights if it is to persuade them to lend to it at all. Although this is far from a complete explanation of firms' decisions about how to finance themselves, it is an element. We will see more on this topic in Chapters 14 and 15.

OIL AND GAS TAX-SHELTER PROGRAMS[11] In the United States in the early 1980s, many oil and gas exploration and development operations were organized through **limited partnerships**, which are hybrid contractual arrangements mixing elements of the forms of both corporations and partnerships. There are two classes of partners in a limited partnership: the *general partners* and the *limited partners*. The limited partners are in a position very like that of the shareholders in a public corporation. They take no role in managing the partnership. Rather, they simply provide the cash as investors to finance its operations, and they enjoy **limited liability**: Their financial liabilities are limited to the amounts they invest. The general partners are like the partners in a regular partnership. They make all the managerial decisions about the partnership's operations, and they have *unlimited liability* for the partnership's debts: Their personal wealth can be seized by creditors if the partnership defaults on its debts.

The federal tax laws that prevailed in the early 1980s partially accounted for the popularity of this organizational form in oil and gas exploration. The partners could often save on taxes if the limited partners paid all the costs of exploring for oil (which were tax deductible when the costs were incurred), whereas the general partners paid the costs of completing wells in which oil is found (which were "capitalized costs" for tax reporting purposes). The general partners and the limited partners would then share any revenues enjoyed when oil was pumped from producing wells.

This tax avoidance scheme is beset with moral hazard problems that arise from the difference in interests it creates between the general partners and the limited partners. The most fundamental of these results because each bears a different kind of expense and receives only a share of the revenues. If a well is found to have oil, the general partners have to decide whether to bring it to completion so it will produce. If they decide to do so, they bear 100 percent of the cost of completing the well but typically receive only 25 percent of the oil revenues, with the rest going to the limited partners. Suppose that after the exploration costs have been sunk, a well is found to have enough oil that the well-completion costs will be just 50 percent of the resulting revenues, so that the partnership as a whole would profit from completing the well. Despite the fact that completing the well would maximize total value, the general partners would not find it in their individual interests to complete the well: Their 25 percent share of the revenue is not enough to cover their 100 percent share of the cost. Furthermore, it would be very hard for the limited partners, with no role in the management of the partnership and probably no expertise in the oil business, to ensure that their interests are being given proper weight in the general partners' decisions.

A second conflict of interest arises when, as was often the case, the general partners are involved in several exploration efforts at the same time and in the same area but have differing shares in different projects. As an extreme example, suppose the general partners have another exploration project on an adjoining tract that they own outright. In that case, by shifting their drilling on the partnership's tract towards the boundaries of their own tract, the general partners can acquire valuable information about the likelihood of finding oil on different parts of their private holdings, with the cost of that information acquisition being borne by the limited partners. Similar, if less severe, problems arise when the general partner is involved in several limited partnerships but has differing interests in each. The general partner may be led to distort his or her allocation of time, effort, attention, and resources among the partnerships, favoring the ones in which he or she has the greatest interest. Again, it would be very difficult for the limited partners to monitor this sort of behavior.

A third conflict often arises when the general partners or their affiliates sell equipment or services to the limited partnership. The problem is that the general partners have an incentive to overcharge on these transactions because the limited partners pay the bills, but the general partners, who make the decisions, collect the money.

All these conflicts are clearly recognized in the industry, and the prospectuses for the limited partnerships often discuss the incentive problems very clearly and candidly. We return to this example in Chapter 7, where some of the means used to offset the incentive problems are discussed.

## CONTROLLING MORAL HAZARD

In order for a moral hazard problem to arise, three conditions must hold. First, there must be some potential divergence of interests between people. Conflicts of interest will not always arise, nor will they arise on all dimensions: Different individuals' interests may naturally be quite well aligned in particular circumstances. However, conflict will occur often, if only because scarcity of resources means that what one person gets another cannot have. Second, there must be some basis for gainful exchange or other cooperation between the individuals—some reason to agree and transact—that activates the divergent interests. Up to this point, simple market arrangements would work: Divergent interests are a factor in almost all exchanges, and yet exchanges are often made successfully without being troubled by moral hazard. The critical third requirement is that there must be difficulties in determining whether in fact the terms of the agreement have been followed and in enforcing the contract terms. These difficulties often arise because monitoring actions or verifying reported information is costly or impossible. However, they could also arise even when both parties know that the contract has been violated but this fact cannot be verified by third parties (such as a court or arbitrator) who would have enforcement powers. This means that the normal market solution will be problematic, because the parties will not be able to write enforceable contracts covering all the crucial elements of the transaction. These three conditions suggest ways to deal with the moral hazard problem.

[11] This section is based on Mark Wolfson, "Empirical Evidence of Incentive Problems and Their Mitigation in Oil and Gas Tax Shelter Programs," in *Principals and Agents: The Structure of Business*, J. Pratt and R. Zeckhauser, eds., (Boston: Harvard Business School Press, 1985), 101–25.

## Monitoring

The first remedy is suggested by the third condition: Increase the resources devoted to monitoring and verification. Sometimes the idea is to prevent inappropriate behavior directly by catching it before it occurs. For example, U.S. corporations are not allowed to publish financial statements until they have been verified by independent auditors, prospectuses describing investments for which funds are sought from the public must be approved by the Securities and Exchange Commission, and health care insurers may have patients obtain a second opinion on a physician's recommendation for some expensive treatment if they are concerned that the treatments may be unnecessary. In other situations, monitoring is intended to decrease the probability of getting away undetected with the socially inefficient, self-interested behavior. In this case, the results of monitoring are the basis for rewards or penalties. For example, workers are often required to punch a time clock, and their pay is reduced or other punishments are imposed if they arrive late or quit early. Monitoring may also be used to support a system of rewards for good behavior.

The payment of cash rewards is itself sometimes subject to a moral hazard problem of reneging. The party who is supposed to pay the reward may misrepresent the outcome of the monitoring, claiming that the other person's behavior was not appropriate and no reward is due. This is likely to be especially easy when the criteria for judging performance are hard to describe or measure precisely, so that evaluations will tend to be subjective. Sometimes, the need to maintain a good reputation is enough to control this temptation. (Reputation effects are discussed in more detail in Chapter 8.) In other circumstances, the efficacy of monitoring may depend on generating evidence verifiable to a court that can enforce payment of the agreed rewards.

A more subtle but related commitment problem arises when punishment is due but carrying out the punishment is costly for the party who is supposed to do it. For example, if company policy requires that an employee who breaks certain rules must be fired, and a valued, hard-to-replace employee is caught in a minor violation of the rules, then the employer may be loathe to carry out the punishment and lose the employee's services. Of course, if the worker foresees that the firm will be unwilling to punish transgressions, the threat of punishment is empty.

COMPETING SOURCES OF INFORMATION Although monitoring requires developing sources of information about the agent's truthfulness and performance, this does not always require direct expenditures of resources. One possibility is to rely on competition among different parties with conflicting interests to develop the needed information. In everyday life, competing sellers will often happily compare the relative merits of their own products against comparative defects in the competing product which the other seller would be unlikely to emphasize. The same phenomenon can occur within organizations, as for example when the navy and air force vie for responsibility for some military mission, each emphasizing its own advantages compared to its competitor. The danger with relying on competing information providers is greatest when they have some common interests that are in opposition to the decision maker's. For example, neither of two sellers of asbestos insulation was likely to emphasize the health hazards of asbestos before these became widely known.

MONITORING BY MARKETS Managerial moral hazard is frequently alleviated by monitoring provided for free by markets. Managers of firms in reasonably competitive product or input markets who do a poor job of generating profits will face a greater probability of failure. The fear of unemployment and of carrying a reputation for having led a firm into bankruptcy may then provide managerial incentives. Similarly, the "market for corporate control" provides incentives by threatening bad corporate managers with loss of their jobs following a takeover or a successful **proxy fight**.[12]

## Explicit Incentive Contracts

In some situations, monitoring actual behavior or the veracity of reports may be simply too expensive to be worthwhile. As we mentioned earlier, however, it may still be possible to observe *outcomes* and to provide incentives for good behavior through rewarding good outcomes. For example, even if it is impossible to monitor the care and skill exerted by machine maintenance personnel, it may still be possible to measure the percentage of time that machines break down. In fact, if the breakdown rate of machines were completely determined by the performance of the maintenance worker, basing pay on that rate would be a perfect substitute for basing it on care and effort. The same would be true even if other factors (such as the machines' inherent quality, the intensity and nature of their use, and the care exerted by the machine operators) also influence the breakdown rate, provided it were possible to control precisely for the effects of these other determinants of breakdown.

Unfortunately, perfect connections between unobservable actions and observed resulting outcomes are rare. More often people's behavior only partially determines outcomes, and it is impossible to isolate the effect of their behavior precisely. For example, a firm's total sales depend not only on the efforts of the sales force but also on a host of other factors: the price and advertising policy of the firm, competitors' prices and promotions, and other conditions that affect customers' demands. Rewarding on the basis of results therefore makes the salespeople's incomes dependent on random and uncontrollable factors. A similar effect arises when the outcomes *are* fully determined by the person's effort but are not measured precisely, instead being only estimated or measured with some unknown, random error. Again, incomes become subject to random variations.

THE PROBLEM OF RISK-BEARING Most people dislike having their incomes dependent on random factors. They are **risk averse**, and would rather have a smaller income whose magnitude is certain than an uncertain income that is somewhat larger on average but is subject to unpredictable and uncontrollable variability. The risks created by incentive contracts are costly to these people. They are not as well off with a risky income as they would be receiving the same expected level of pay for certain, and they thus have to be paid more on average to convince them to accept these risks. From the employer's perspective, this extra income is a cost of using incentive pay.

Moreover, this cost can be a real one to society, one that can reduce overall efficiency. The employer often is more tolerant of risk and better able to bear it than are employees. In the extreme case, where the employer is a well-financed and widely held corporation whose stockholders keep their wealth in broadly diversified portfolios, the stockholders can be assumed to be *risk neutral*—concerned mostly with expected returns and virtually indifferent about variability in the net earnings of the firm, especially variations of the magnitude of an individual worker's performance pay. Tying workers' pay to their job performance means that a source of the variability of earnings is transferred from the owners to the workers: When things go well on the job, some of the extra returns accrue to the workers, and when things go badly, the

[12] The stockholders in a corporation have the right to elect the directors and to vote on certain major policy decisions at stockholders' meetings. Few stockholders ever attend these meetings, however. To allow for this, they are permitted to give their *proxy* to someone else to cast their votes on their behalf. Typically, the proxy is given to management. In a proxy contest, rival groups will attempt to win stockholders' proxies so that they can elect different directors or prevent management and the current directors from enacting a policy change that the group opposes. See Chapter 15.

impact on the owners is cushioned by the lower levels of incentive pay. However, transferring risk from the owners (who care little about the risk and benefit little from its reduction) to the workers (who may strongly dislike bearing risk) means that the total costs of the given amount of risk in the system are increased.

RISK COSTS AND INCENTIVE BENEFITS Designing efficient incentive contracts involves balancing the costs of risk bearing against the benefits of improved incentives. Insulating risk-averse employees' pay from variations in measured job performance minimizes the costs of risk bearing, but it also eliminates monetary performance incentives. Shifting risk to the employees strengthens their incentives because their pay now depends on actual performance, but the costs of risk bearing rise as well. The efficiency principle suggests that observed contracts will tend to be efficient, subject to the constraints imposed by observability problems.

One implication of this analysis is that it is inefficient to use contracts that make risk-averse employees bear avoidable risks unless the contracts also provide useful incentives.[13] For example, consider a firm whose risk-neutral owners want to maximize profits and which has a problem motivating production workers to be productive. Because the owners care about profits, one possibility is to provide incentives for everyone by paying bonuses based on profitability. This exposes workers to income variations arising not just from their own productivity, however, but also from all the other factors influencing profits that are beyond their control: input prices and availability, the efforts of the sales force, the quality of executive decisions, variations in demand and in the interest rate the firm has to pay on its debts, the actions of competitors, and so on. In this case, it may be preferable to use an incentive plan based not on profits but on direct measures of the contributions made by individual workers or work group, such as the volume of output, the number of defects, the number of days absent from work, and so on. Even these measures expose the workers to risk, because productivity is not completely under their control, but they do insulate them from some unnecessary risks.

The basic idea behind incentive contracts is that of achieving **goal congruence**: An appropriately designed reward system causes self-interested behavior to approximate the behavior the designer wants. Alternatively, we can think of a well-designed incentive scheme as removing the conflict of interests by effectively altering individual objectives, aligning them more closely with those of the designer. We will usually think of incentives as altering rewards to increase the benefits associated with the desired behavior; for example, motivating employees' interest in profit seeking by tying their pay to profitability. However, behavior can also be modified through job design, employee involvement programs, and the provision of a better work environment, all of which reduce the unpleasantness of work and lower the costs to employees of providing effort. Requiring office workers to be at their desks during certain hours can be seen in similar terms. Because they have to be at the office, they may as well do their jobs, although if they were free to be elsewhere, they would find other things to occupy their time.

We give a (relatively) simple mathematical example of what is conceptually involved in designing an efficient incentive contract in the appendix to this chapter. In Chapter 7 we examine this issue in much more detail and develop a number of principles that can be used to understand and evaluate actual contracts and to guide contract design. Also, in Chapter 12 we examine managerial issues that arise in using incentive pay in organizations.

[13] The point is closely related to the adage that people should be held responsible only for things under their control. Actually, the adage with this phrasing is misleading, as seen in Chapter 7.

## Bonding

In some industries, it is common to require the posting of bonds to guarantee performance. The bond is a sum of money that is forfeited in the event that inappropriate behavior is detected. For example, contractors often must post a bond that they lose if the project is not completed by the agreed date and in the agreed manner. Similarly, the capital provided by the owners of a bank or an S&L acts like a bond because in the event of losses the capital must be paid out to meet obligations. In the early 1970s, Electronic Data Systems Corporation (EDS)—Ross Perot's computer service company that was later acquired by General Motors—required trainees who resigned within three years of joining the firm to pay the firm $12,000.[14] This bonding was designed to prevent employees from receiving costly training without doing substantial work for the firm. The $12,000 amount was comparable to an engineer's annual salary at the time.

Posting a bond can be a very effective way to provide incentives, but the problem is that people often will lack the financial resources to post a sufficiently large bond. This is especially the case when the gains from cheating are large and the probability of getting caught is small, so that the bond would have to be large to give an adequate incentive. These ideas are examined more carefully in Chapter 8, but one application that sheds light on the puzzle of positively sloped age/wage profiles can be discussed here.

AGE/WAGE PATTERNS, SENIORITY PROVISIONS, AND MANDATORY RETIREMENT As noted in Chapter 5, pay tends to increase with age and experience, even after controlling for productivity. In Chapter 5 we offer an explanation for this pattern based on inducing self-selection to reduce employee turnover. Bonding as a deterrent to employee shirking has been suggested as an alternative explanation by Edward Lazear.

Suppose that the firm can fire any workers detected shirking. We may think of workers who shirk as receiving some valuable benefit, such as a reduced level of stress or more time to pursue personal interests, which cannot be taken away from them. If workers were to post bonds of sufficiently greater value than these benefits, and if being caught cheating resulted in losing the bond, then they would not cheat. Their value to the firm would be increased by the bond and, with competition among employers, so too would be the amount they would earn. In any case, when it is efficient for workers not to cheat and shirk, the bond may allow efficiency to be achieved. If the gain from cheating is substantial, however, or the likelihood of getting caught is small, workers may not be able to afford to post a big enough bond, and the potential efficiency gain would be lost.

Suppose the firm in this circumstance makes a credible promise to the workers that, late in their careers, it will pay them more than the value of what they produce and thus of what they could earn elsewhere. If the firm pays workers less than their marginal products early in their careers, then the value of lifetime earnings and the firm's total outlay need not be affected by this scheme. As the years of high pay draw near, however, the high promised wages serve as a bond that the worker would forfeit by dishonest behavior or shirking: The wage pattern duplicates the effect of a bond. Therefore, the observed pattern of wages might be explained by a need to make workers value their jobs in order to ensure honest, hard-working behavior.

Strikingly, a mandatory retirement provision will be necessary for efficiency under this scheme. For efficiency, people should retire when the value of what they produce just equals the private cost to them of continuing working. If they were paid

[14] Doron Levin, *Irreconcilable Differences: Ross Perot versus General Motors* (New York: Plume, 1989), p. 46.

their marginal products, they would choose to retire at the efficient date, when the extra income just balances the increasingly high costs of continuing work. With wages late in life exceeding marginal productivity, however, some people will want to continue working too long because their pay exceeds the social value of their output. They will not retire voluntarily at the appropriate date. Thus, mandatory retirement is necessary for efficiency. Furthermore, at the start of their careers, workers would be happy to sign contracts agreeing to a mandatory retirement date, even though once it arrives they will be unhappy about being forced to stop working and earning.

This scheme also necessitates some sort of mechanism to make the firm's promise credible. The danger is that the firm will renege on its agreement by letting senior workers go once their pay exceeds their current productivity. Again, a concern with reputation may work here, but it is perhaps less likely to be effective when workers have few other employment options and so the incentives to avoid a firm that has cheated are weak. In this case, a seniority rule in layoffs can play a useful role. If the firm wants to lay off a senior worker who is earning a lot, it must first lay off all the more junior people whom it is paying less than they are worth.

## Do-It-Yourself, Ownership Changes, and Organizational Redesign

Moral hazard in agency settings can sometimes be overcome by eliminating the agent and having the principals act on their own behalf. This is often impossible, however—you cannot very well be your own surgeon, for example—and in any case it sacrifices the gains of specialization.

In market settings, changing ownership patterns to bring the affected transactions within a single organization can help overcome some moral hazard problems. (We have already noted in Chapter 5 that unified ownership can be a response to inefficiencies arising from bounded rationality and private information.) For example, if a firm and a supplier are frequently involved in complex transactions that are marked by such manifestations of moral hazard as possible cheating on quality, it may be efficient for one firm to acquire the other. In that case, the differing interests (each firm's own profit) become merged, and many of the incentive problems are overcome. A simple mathematical example illustrates how this can be effective.

INCENTIVES AND OWNERSHIP PATTERNS: AN EXAMPLE Suppose that two firms face a joint opportunity that requires both of them to make some investment. Let us denote the measured value of the investments by firms A and B by $M_A$ and $M_B$. These measured amounts need not be the assets' actual values, however, and this is where moral hazard enters because the firms' wills individually decide the actual amounts they will invest. For example, a physical asset might be valued at $100 in the accounts because it cost $200 a year ago and had an estimated useful life of two years. However, the actual value of that asset today might be only $50 because new machines using a new process have made the old process obsolete. Similarly, the value of a year of an employee's time might be recorded at an amount equal to his or her wage, but the wage might not accurately reflect the employee's value to the company. The employee might be a young engineer, straight out of school, who is expected to go through a period of learning and low productivity (when one subtracts the cost of mistakes). Or, the employee might be an up-and-coming executive who has successfully managed other special projects and is more valuable to the company than others of the same rank.

Let $V_A$ and $V_B$ be the *actual* value to the two companies of the resources they invest in the venture. The expected revenue from the venture is assumed to be 1.5 times the actual value of the total investment minus $600, that is, $1.5 \times (V_A + V_B) - 600$. Higher real investments lead to more revenue for the project on average.

**Table 6.4 Company A's Profit Calculation**

| | |
|---|---|
| B's investment | $V_B$ |
| A's investment | $V_A$ |
| Total investment | $V_A + V_B$ |
| Total expected revenue | $1.5(V_A + V_B) - 600$ |
| A's expected revenue (a one-half share) | $0.75 \times (V_A + V_B) - 300$ |
| A's net expected profit (revenue minus investment) | $0.75 \times V_B - .25 \times V_A - 300$ |
| A's profit-maximizing choice | $V_A = 500$ |
| B's expected choice | $V_B = 500$ |
| A's expected *loss* | $50 |

Suppose the two companies write a contract according to which each will invest $1,000 in the project. This can only refer to measured investments, because even if the companies were able to distinguish the values of the assets being used, there would be no objective way for a court to verify the level of unmeasured investment in order to enforce the agreement. As equal partners, the companies agree to divide the revenues generated by the venture equally. If the companies were to invest so that measured and actual values were the same, that is, so that $M_A = M_B = V_A = V_B = \$1{,}000$, then the investment would be profitable for both companies. The total revenues generated by the project would be $1.5 \times (V_A + V_B) - 600 = \$2{,}400$, which is more than the total investment of $2,000. Each company would expect to receive $1,200 in revenues from an investment of $1,000, for a net profit of $200.

Now, suppose that each company is free to choose its investment in the project so that, even though the measured value is exactly $1,000, the actual value can be anything between $500 and $1,500. What choices will the companies make? Will the venture be a profitable one?

Let's look at the matter from company A's perspective. Whatever amount $V_B$ company B actually invests, company A's expected profits will be its share of the total revenues minus its investment (see Table 6.4).

According to Table 6.4, each extra dollar of value that firm A invests in the venture yields $1.50 in extra revenue, but A's share of that $1.50 is just $0.75. Therefore, A loses $0.25 on each extra dollar of investment. A's profit-maximizing choice is to make the minimum investment of $500, as indicated in the seventh line of the table. If A understands this calculation (and expects that B does too), then it will expect B to invest only $500 as well. In that case, the total expected revenues of the venture will be just $900, whereas the required investment will be $1,000. Each firm will expect to *lose* $50, so the investment will not be made. An apparently profitable business opportunity will be lost because A expects B to take a free ride on their investment and B expects A to do likewise.

Of course, if the two firms were to merge, and the decisions about investment were brought under a single individual who pays the entire costs and collects the full benefits, then the problem described here would be eliminated.

THE DETERMINANTS OF OWNERSHIP As our examples related to moral hazard in employment have shown, a transaction that is "integrated" or brought "in house" does

not automatically align incentives. The problem is that integration only transforms a self-interested manager who formerly worked for the supplier into a self-interested manager who works for the firm. The basic incentive problem may still need to be solved.

The upshot is that merger does not always eliminate the incentive problem that exists between separate firms. Compounding the problem is the fact that there are additional unavoidable costs to bringing previously separate activities under common direction. An important component of these are the *influence costs*, which increase with the increased potential for central control of activities in an integrated organization. Although the *influence activities* that give rise to influence costs are a form of moral hazard, they are of such importance in understanding organizations that they deserve separate treatment.

## Influence Activities and Unified Ownership

What costs are involved in bringing two separate organizations under unified direction? Why can't the merged entity do everything the separate components did and more? What are the limits, if any, on the efficient size of organizations? Why isn't all economic activity organized in a single firm?

From our discussion in Chapter 2, it is clear that the answer must be that bringing everything within a single organization involves inefficiently high transactions costs of some sort. But what are they? In fact, until recently, little attention has been given to the task of identifying the transactions costs of internal, nonmarket organization. This is a subtle matter. In actual organizations, much time and ingenuity is spent overcoming transactions costs: Witness the example of the development of the multidivisional form in Chapter 1. Moreover, a strikingly simple idea—the policy of selective intervention—undercuts many of the possible candidates that come to mind as distinctive disabilities of unified control.

### Unified Ownership and Selective Intervention

Suppose it is efficient for two parts of a big organization to be independent and operate as separate entities. Then, in the original organization the center could direct the two units to conduct their transactions at arm's length as if they were not both part of a single structure. For example, when market transacting works well, why not replicate its operation within the firm, using internal, transfer pricing? Meanwhile, where there are efficiency gains to be had from deviating from the patterns of transactions that would occur in the market, why not have the central management *selectively intervene* in the operations of the component units to ensure that the gains are realized?

Following a policy of selective intervention consistently ought to mean that the unified organization can do everything the separate pieces could do, and do so at least as well. There would then be no bound on the efficient size of the organization. Why then is all activity not brought under a single firm? The logical answer must be that adhering to a thoroughgoing policy of selective intervention is impossible. But why should this policy be infeasible? **Influence activities** provide part of an answer.[15]

Influence activities arise in organizations when organizational decisions affect the distribution of wealth or other benefits among members or constituent groups of the organization and, in pursuit of their selfish interests, the affected individuals or groups attempt to influence the decision to their benefit. The costs of these influence activities are **influence costs**.

[15] A broader discussion of influence costs is found in Chapter 8.

### Influencing Interventions

The fundamental difficulty with the policy of selective intervention is that it requires that there be a decision maker with the *power* to intervene who *collects information* with which to make decisions: These things can by themselves impose costs on the organization. The most obvious costs are the decision maker's salary and the cost of providing information to support the decision-making system, including the time that lower-level decision makers spend reporting information to the decision maker. Often more important is that individuals and units within the organization may have selfish reasons to seek unproductive interventions, and they may expend resources trying to *influence* the decision maker to bring them about. Even when the attempts fail, the resources expended in these influence activities represent a cost that brings no offsetting gain. When they do succeed in influencing the central decision maker to intervene inappropriately, there are further costs in bad decisions being made and implemented. Finally, if the organization recognizes these possibilities and adjusts its structure, governance, policies, and procedures to control attempts at influence, these deviations bring further costs. All of these are elements of influence costs.

As is evident, the magnitude of influence costs depends on the existence of a central authority, the kinds of procedures that govern decision making, and the degree of homogeneity or conflict in the interests of organization members. All this is treated in more detail in Chapter 8. Here, we focus on how influence costs limit the optimal scope of formal organizations.

When two previously separate organizations are brought under a common, central management with the power to intervene, the scope for influence increases and influence costs increase. For example, members of one unit can try to influence top management to transfer resources from the other unit to theirs. They can argue that they have better investment opportunities and so can better use the funds being generated in the other division, or that they have more valued uses for the most talented people now assigned to the other group, or that all marketing, or production, or research and development (R&D) should be consolidated in a single unit (theirs!) rather than remaining inefficiently spread over several units. The other group will have a similar incentive to defend itself and even to counterattack. It can argue that other units should be required to purchase its outputs, even though the outside market may provide superior or cheaper substitutes, because doing so helps cover corporate overhead or helps build the firm's core competencies, or it may complain that equity, morale, and ultimately productivity demand that its members be paid as well as those in another group whose members may be especially productive or may have particularly valuable skills or knowledge. Large amounts of time, ingenuity, and effort may go into these attempts at influence, and huge amounts of the central executives' time can be consumed dealing with them.

Of course, none of this would occur if there were no central authority with the power to make the proposed changes. Thus, although the merged organization may be able to achieve things that were not possible before, it also suffers costs that were not present when the parts were separate.

### Influence Costs and Failed Mergers

This logic gives insight into the great frequency with which corporate mergers and acquisitions apparently fail. In a study of the diversification records of 33 large U.S. corporations between 1950 and 1986, Michael Porter found that fully 60 percent of the acquisitions in new fields of business by these firms were later divested, and 61

percent of the firms ended up divesting more of their acquisitions than they kept.[16] Although not all divestitures of previous acquisitions necessarily represent failures, even the sophisticated firms in Porter's sample had real problems making acquisitions work.

There are obviously major problems involved in attempting to integrate two different organizations with their own unique histories, their own ways of doing things, their own reporting and control systems, their own pay and benefit schemes, and so on. To focus on a pure case in which these factors should be of minimal concern, consider a pure conglomerate merger in which one firm acquires another with the intent of running it as a completely separate division, intervening in its operations only when there are clear gains to doing so. Even here, influence costs present problems that may cause the merger to fail.

TENNECO'S ACQUISITION OF HOUSTON OIL AND MINERALS A well-documented example is the 1980 acquisition of Houston Oil and Minerals Corporation by Tenneco, Inc., which was then the largest conglomerate in the United States.[17] Houston's business was finding, developing, and bringing petroleum and mineral deposits into production. The company was very aggressive and quite successful before its acquisition by Tenneco. Tenneco's stated intent was to run Houston as a separate company, maintaining the entrepreneurial, risk-taking style that had marked it as an independent concern. In particular, it planned to maintain a separate compensation and reward system at Houston that would provide unusually large individual payoffs to professionals for successful discovery and development of petroleum reserves. (Several Houston explorationists had become wealthy with the bonuses they earned from successful explorations, and such packages were common among smaller firms in the industry.) Yet Tenneco had great difficulty developing such a plan, and it ultimately failed to do so. Within a year, more than a third of Houston's managers, a quarter of its exploration staff, and almost a fifth of its production people had left the company for better opportunities elsewhere. This severely hampered operations, and ultimately it became impossible to maintain Houston as a distinct unit within the firm.

Tenneco's costly failure to institute the intended reward policy apparently resulted from a concern for equity in pay across the organization. The Tenneco Vice President of Administration was quoted in *The Wall Street Journal* as saying: "We have to ensure internal equity and apply the same standards of compensation to everyone." Meeting this perceived need was very costly. It contributed to the exodus and to the ultimate failure of the acquisition. Failure to meet this need might well have been even more costly, however. Tenneco's 100,000 employees, jealously looking at the huge bonuses that would have been paid to the few hundred Houston professionals, might have consumed large chunks of their superiors' time with their jealous complaining and their attempts to get some of these funds for themselves. Given the relative sizes of the two groups, the overall impact on productivity at Tenneco could have been disastrous.

---

[16] Michael Porter, "From Competitive Advantage to Corporate Strategy," *Harvard Business Review* (May–June 1987), 43–59.

[17] This example is discussed by Oliver Williamson in *The Economic Institutions of Capitalism* (New York: The Free Press, 1985), p. 158. The primary source is George Getschow, "Loss of Expert Talent Impedes Oil Finding by New Tenneco Unit," *The Wall Street Journal* (February 9, 1982), A-1. The quotation in the next paragraph is from this story.

## SUMMARY

The term *moral hazard* originated in the insurance industry, where it referred to the tendency of people who purchase insurance to alter their behavior in ways that are costly to the insurance company, such as taking less care to prevent a loss from occurring. Within economics, the term has come to refer to any behavior under a contract that is inefficient, arises from the differing interests of the contracting parties, and persists only because one party to the contract cannot tell for sure whether the other is honoring the contract terms. Moral hazard problems arise frequently in *principal-agent* relationships, where one party (the "agent") is called upon to act on behalf of another (the "principal"), because the agent's interests commonly differ from the principal's and the principal cannot evaluate how well the agent has worked or whether the agent has been honest.

The savings and loan crisis in the United States illustrates the problem of moral hazard and how it is most often dealt with in ordinary business transactions. The difference of interests between the owners of the S&L and the federal insurance agency (the FSLIC) arose because the owners benefited from risky investments when they turned out well, but the costs of failures are borne by the insurance agency. When the agency failed to monitor and control the S&L, this led the S&L management to make risky investments or even to engage in fraud in a way that was costly to taxpayers. Competition for depositors' funds only intensified this effect, increasing the interest rates paid on deposits and forcing more conservative S&L management to find higher-yielding—and hence usually riskier—investments.

The relation of depositors to their S&L is similar to the relationship of lenders to any other kind of firm. Normally, lenders protect their money by imposing controls and requiring reporting by and audits of the borrower. What distinguished the S&L case is that the depositors, being insured, had little reason to monitor the savings institutions, and the federal government, whose money was at risk, did not monitor on its own behalf, in part because powerful congressmen were protecting the S&Ls.

Problems of fraud and excessive risk taking similar to the S&L problem can be found in many federally insured programs. Among those described are programs insuring workers' retirement benefits, farmers' crops, mortgage loans, and student loans. Similar problems exist in private-sector insurance programs, but these tend to be less severe partly because profit-oriented insurers monitor the insured more carefully and partly because private insurers refuse to offer insurance when the moral hazard problem is too severe.

Moral hazard is not only a problem of markets, but exists in other kinds of organizations as well. Air traffic controllers seeking to collect disability benefits have "punched out," causing incidents that seemed to indicate that they suffered job-related stress and were unable to continue their duties safely. The general partners in oil and gas drilling partnerships sometimes fail to complete wells that are profitable for the partnership as a whole because their own interests differ.

Various means are available to control the moral hazard problem. One is explicit *monitoring*, which can reduce the information problem that is a fundamental component of moral hazard. A second is the use of *incentive contracts* that pay for output performance when inputs cannot be measured. *Posting a bond* that is forfeited if the agent is caught cheating can be effective in a principal-agent relationship. This bond can be implicit in the rising pattern of wages over a worker's career: A worker caught cheating after several years of employment stands to lose the high wages paid to more senior workers. Sometimes, the whole problem can be avoided by the "do-

it-yourself" solution, which does not rely on an agent with differing interests. Similarly, a firm can sometimes eliminate conflicts of interest with its suppliers by integrating vertically, though this does not eliminate any individual differences of interests between the formerly independent manager of the supplier and the same person who is now an employee of the firm.

An especially important category of moral hazard is the category of *influence activities* and the associated costs, known as *influence costs*. These arise when employees divert effort to influence organizational decisions. Even if those decisions are not ultimately affected, the time, effort, and ingenuity devoted to attempts at influence are unavailable for more productive activities. Influence costs are one of the important costs of centralized control and help to explain the importance of organizational boundaries. These costs are largely eliminated when there is no decision maker with authority to make the decisions that employees wish to influence, and this condition can sometimes be brought about by creating legal or other boundaries between operating units.

## BIBLIOGRAPHIC NOTES

As with so much else in the economics of organizations, the problems of misaligned incentives and what we now call moral hazard were noted and understood by Adam Smith: see his discussion in *The Wealth of Nations* of the incentives in joint stock companies (Book V, Chapter I, Part III, Article I) and of the incentives for university teachers (Book V, Chapter I, Part III, Article II). The nature of moral hazard and its importance to economic analysis was made explicit by Mark Pauly in the context of an article on the economics of health insurance by Kenneth Arrow. The principal-agent model, which underlies much of the discussion in this chapter, has several early contributors, including James Mirrlees, Michael Spence and Richard Zeckhauser, and Steven Ross. More recent references are given in the next chapter. The explanation of the age/wage profile and mandatory retirement in terms of bonding is due to Edward Lazear. A useful discussion and development of Lazear's bonding model is given by Lorne Carmichael.

Adolphe Berle and Gardner Means began the debate about whether the ownership structure of the modern corporation has made it particularly susceptible to managerial moral hazard. The papers published in the *Journal of Law and Economics* [26 (June 1983)] from the Hoover Institution conference on "Corporations and Private Property" held to commemorate the fiftieth anniversary of the publication of the Berle and Means book give some flavor of current thinking on this issue, which in turn has been central to the debate over hostile takeovers. The "Symposium on Takeovers" in the *Journal of Economic Perspectives* [2(1988), 3-82] provides an overview of economists' views on this debate, which we consider in Chapter 15.

The modeling of bankruptcy as a means for effecting efficiency-enhancing changes in control is due to Phillippe Aghion and Patrick Bolton. The determinants of firms' financing decisions are treated in Chapters 14 and 15, and further references are given there.

The importance of the policy of selective intervention was emphasized by Oliver Williamson. The concept of influence costs as a major element of the transactions costs of nonmarket organization was developed by the present authors. The theory of information provided by competing sources is developed in our *Rand Journal of Economics* paper.

## REFERENCES


Aghion, P., and P. Bolton. "An 'Incomplete Contract' Approach to Bankruptcy and the Financial Structure of the Firm," *IMSSS Technical Report*, No. 536 (Stanford, Ca: Stanford University, 1988).

Arrow, K.J. "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53 (1963), 941–73.

Berle, A., and G. Means. *The Modern Corporation and Private Property* (New York: MacMillan, 1932).

Carmichael, L. "Self-Enforcing Contracts, Shirking and Life Cycle Incentives," *Journal of Economic Prespectives*, 3 (1989), 65–83.

Lazear, E. "Why Is There Mandatory Retirement?" *Journal of Political Economy*, 87 (December 1979), 1261–84.

Milgrom, P., and J. Roberts. "Relying on the Information of Interested Parties," *Rand Journal of Economics*, 17 (1986), 18–32.

Milgrom, P., and J. Roberts. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity," in *Perspectives on Positive Political Economy*, J. Alt and K. Shepsle, eds. (Cambridge: Cambridge University Press, 1990).

Mirrlees, J. "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 38 (1971), 175–208.

Pauly, M. "The Economics of Moral Hazard," *American Economic Review*, 58 (1968), 31–58.

Ross, S. "The Economic Theory of Agency: The Principal's Problem," *American Economic Review*, 63 (1973), 134–39.

Spence, A.M., and R. Zeckhauser. "Insurance, Information and Individual Action," *American Economic Review*, 61 (1971), 380–87.

Williamson, O. *The Economic Institutions of Capitalism* (New York: The Free Press, 1985).


## EXERCISES

### Food for Thought

1. Widespread fraud brought down some S&Ls. Does deposit insurance itself make fraud more attractive? How did the changes in regulation of the S&Ls contribute to the problem of fraud?

2. As part of a plan to rescue the savings and loan industry, in late 1988 the U.S. government encouraged private investors to purchase the assets of failing savings and loans. They offered guarantees of principal and, in some cases, interest on some of the properties taken over by the S&Ls when borrowers defaulted on loans. What effect would you expect these guarantees to have on the behavior of the new owners?

3. In automobile collision insurance and health insurance, the insurance policy often has a provision calling for a *deductible* according to which the portion of any insured loss up to some fixed limit, such as $500, is paid for by the insured person; only the excess is paid for by the insurance company. In addition, health-insurance policies often provide for *copayments* by the insured, according to which the insurance company pays only some fraction, such as 80 or 90 percent, of the medical costs in excess of the deductible, until the insured has paid some maximum amount (such as $2,000 in a year). What economic function do deductibles and

copayment provisions serve? Why do we see deductibles but not copayments used in automobile insurance policies?

4. Farmland on which annual crops are grown is often rented, but orchards and other perennial crops are more often grown by the owners of the land. How can this be explained?

5. In a possibly apocryphal story, a nineteenth-century English traveler in China was shocked that the oarsmen rowing the boat in which the traveler was riding were brutally whipped by a ferocious overseer if they slacked off on their rowing. The traveler was even more shocked (so the story goes) to learn that the oarsmen owned the boat and hired the overseer to beat them! How would you explain this?

6. We attribute the prevalence of limited partnerships in oil and gas exploration to the tax advantages of having some claimants pay some of the costs and other claimants pay other cost elements. But the tax advantages might have been achieved by a regular partnership in which different partners had different forms of claims. What are the advantages then of the limited partnership form of organization?

7. Stanford University's Honor Code forbids faculty from monitoring students during examinations. What effect would you expect this to have on student behavior? On the relationships between students and faculty?

## Quantitative Problems

1. Suppose there are two firms, Firm A and Firm B, that are considering making a joint investment in R&D. The total payoff from the project is $200 \times (V_A + V_B)^{1/2}$, where $V_A$ and $V_B$ are the values of the two investments. The two firms expect to share this payoff equally while each absorbs the cost $V_A$ or $V_B$, of its investment. Show that the value-maximizing plans are those where the total investment of the two firms is $10,000. How much total value is created in this way?

2. In question 1, suppose now that the firms cannot enforce a contract specifying levels of investment for each, because they cannot observe the real value of the investments that are made. Show that if Firm A expects Firm B to invest $V_B$, it can do no better than to invest $2,500 - V_B$. How much, then, will be invested in total? How much total value is created?

3. Suppose that if the firms do not sign a contract, that they can develop their own versions of the research project in competition with one another. A marketable product will result for either firm provided it spends at least 35 percent as much as its competitor. If $V_A \geq .35 \times V_B$, then Firm A's net profit will be $200 \times (V_A - .35 \times V_B)^{1/2} - V_A$, and correspondingly for B. Show that if each firm expects the other to invest 10000/.65, then it will choose to invest an equal amount. What will the total profits be? Would you expect the firms to reach a joint venture agreement under these circumstances?

4. (*Forcing contracts*). In principal-agent problems of efforts provision, moral hazard may not be a problem if the structure of uncertainty allows the principal to infer precisely whether the agent has failed to perform as desired. To see this, suppose in the context of the example in the Appendix that the matrix relating the probabilities of various outcomes were changed so that the high level of revenues (30) was sure to occur if the worker supplies the high level of effort ($e = 2$), but either level of revenue could still happen if the low level of effort ($e = 1$) is provided. Thus, the first row of Table 6.5, corresponding to $e = 1$, is unchanged, but the entries corresponding to $e = 2$ become 0 and 1 instead of 1/3 and 2/3.Rewrite the incentive and participation constraints and show that it is possible to design a *forcing* contract that motivates the worker to work hard, supplying $e = 2$, without placing any risk on him or her. How much is the worker paid if revenues of 10 are realized? How much when revenues are 30? What are the expected utilities of the two parties? Is there any cost in this case to effort not being observable, that is, could the parties do better if effort were observed? Would it be possible to achieve this sort of result if the low level of effort surely resulted in revenues of 10, but the high level of effort could result in either revenues of 10 (with probability 1/3) or 30 (with probability 2/3)? Why or why not?

5. (*Selling the firm to a risk-neutral agent*). In many principal-agent problems of effort provision, moral hazard is costly if the agent is risk-averse because making his or her pay reflect the full marginal impact of his or her effort choices imposes costs on the agent that could be avoided if the risk-neutral principal absorbed the variability in incomes. Again in the context of the example in the Appendix, show that if the agent is risk-neutral, with utility function for income $w$ and effort $e$ of $u(w,e) = w - (e-1)$, then it is possible to acheive the same expected utilities for both parties as would result if effort were observable by having the agent bear all the risk and the principal receive an amount that is independent of the realized level of revenues.

## Appendix: A Mathematical Example of Incentive Contracting*

The purpose of this appendix is to develop a relatively simple example of what is conceptually required to determine an efficient contract in the presence of unobservable effort and the consequent moral hazard problem. In the next chapter we develop these matters much more fully, although with less visible mathematics. The example is a principal-agent problem. As we mentioned earlier, in the standard language of incentive theory, an *agent* is someone who does work on behalf of another person, called the *principal*. We will think of agent as a worker and principal as an employer.

Suppose the principal is risk neutral, caring only about the expected amount of money he or she receives, and the agent is risk averse and also averse to providing more than a minimal amount of effort. In particular, suppose the agent evaluates wage income and effort according to a utility function of the form $U(w,e) = \sqrt{w} - (e - 1)$, where $w$ is the wage and $e$ is the effort level. According to this mathematical formula, the marginal utility of income is $1/(2\sqrt{w})$, which is decreasing in the wage level. As we will see in Chapter 7, this property corresponds to risk aversion. The $(e - 1)$ term is the cost of effort, and its form reflects the idea that providing effort is costly only when effort exceeds 1 unit.

Suppose that two effort levels are possible: $e = 1$ and $e = 2$. The agent also has outside job opportunities, and to get him or her to accept employment the agent must be offered at least as much satisfaction as he or she can get working elsewhere. We model this by imposing a requirement that the job provide the agent with a utility level of at least some expected utility $u$, which we can interpret as the expected utility value of his or her next-best alternative. To keep the arithmetic simple, let us suppose that this minimum acceptable utility level $u$ is 1.

The agent's efforts are assumed to help increase the revenues of the firm. Depending on those efforts, various possible levels of revenue might be received. However, the outcome also depends on random factors that neither the principal nor the agent can observe or control. Table 6.5 gives the probabilities of the possible outcomes for each level of effort. For example, when $e = 1$, Revenue is 10 with probability 2/3 and 30 with probability 1/3.

With $e = 1$, the expected revenues are $(2/3) \times 10 + (1/3) \times 30 = 50/3$, whereas lifting the effort to $e = 2$ raises the expected receipts to $(1/3) \times 10 + (2/3) \times 30 = 70/3$. Effort is productive in raising the probability of the good outcome and thus the expected receipts.

If $e$ were observable and the parties wanted it set at 2, the solution would be for the contract to specify that $e = 2$ and that the agent be paid enough to get him or her to agree to take the job when he or she provides $e = 2$, and to be paid nothing if the agent picks $e = 1$. Because the contract calls for a fixed wage $w$ if the agent provides the required effort, the agent bears no uncontrollable risk; the income will be $w$ for sure. The revenues, of course, remain random, but this risk is borne entirely by the risk-neutral principal. This allocation of risk is efficient. Putting any variability in the agent's pay would necessitate compensating him or her for bearing risk, whereas the risk-neutral principal is indifferent about the risk. The pay needed to get the agent

* Those who are not familiar with the theory of expected utility and decisions under uncertainty might prefer to skip this Appendix until they have studied these topics, which are developed at the start of Chapter 7.

**Table 6.5 Probability of Outcomes For Differing Levels of Effort**

| | Revenue | |
|---|---|---|
| Action | $R = 10$ | $R = 30$ |
| $e = 1$ | $p = 2/3$ | $p = 1/3$ |
| $e = 2$ | $p = 1/3$ | $p = 2/3$ |

to agree to the contract is determined by the utility function and the minimum available elsewhere:

$$\sqrt{w} - (e - 1) = \sqrt{w} - (2 - 1) \geq 1, \quad \text{or} \quad w \geq 4$$

So long as the pay is at least 4, the agent will not prefer to take a job elsewhere. Because the principal has no reason in this model to give the agent any more than necessary, the principal gets an expected return net (of the pay to the agent) of $(70/3) - 4 = (58/3)$.

In contrast, if the principal wants only the low level of effort, this can be achieved at minimum cost by paying 1 in either event. This puts no risk on the agent. It again gives the agent an expected utility of 1, but (47/3) to the principal. Thus, the wage cost of the extra effort is $4 - 1 = 3$, both to the principal and to "society," whereas it raises expected receipts by $(20/3) > 3$. Thus it is worthwhile to require and pay for the higher level of effort.

In any case, if $e$ is observable, then regardless of the desired level of $e$, the efficient contract protects the worker from having to bear any uncontrollable risk while, if the high level of effort is efficient, the contract requires the agent to work at the higher level but compensates him or her for it.

Things are different when only the level of revenues, but not the level of effort $e$, is observable. In this case, the principal cannot effectively insist that the agent take a particular level of effort. Effort is not observable, and revenues are not fully determined by effort, although they are responsive to it. Revenue levels of 10 and 30 are both possible no matter what the agent does, so a bad outcome might be attributable to bad luck rather than shirking, and a good outcome might occur by sheer good luck regardless of what the agent does.

If a high level of effort is desired but the agent is averse to working that hard, the way to motivate high effort is to pay more for a good outcome than for a bad outcome. This requires exposing the agent to some income risk.

If the principal wants $e = 2$, then the agent's expected utility when he or she picks $e = 2$ must exceed that when he or she slacks off and picks $e = 1$. Let $y$ be the amount the agent receives under the incentive contract when the outcome is 10 and $z$ be the pay when the receipts are 30. Then the agent's expected utility in picking $e = 2$ is

$$(1/3)(\sqrt{y} - 1) + (2/3)(\sqrt{z} - 1)$$

(where we have used the probabilities that correspond to the high level of effort in evaluating the expected wage), whereas his or her expected utility from picking $e = 1$ is

$$(2/3)(\sqrt{y} - 0) + (1/3)(\sqrt{z} - 0)$$

In this expression we used the probabilities corresponding to $e = 1$. For the agent to

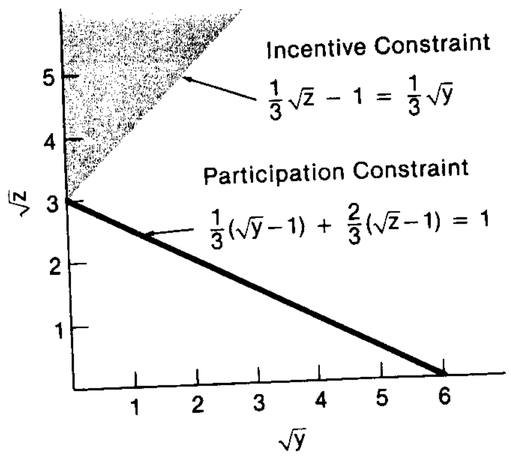


**Figure 6.1:** In this example, the point (0,3) satisfies both the incentive and participation constraints at least cost to the principal. Note that the graph is expressed in terms of the *square roots of y and z*.

be willing to pick the higher effort level, the first of these expressions must be at least as large as the second:

$$(1/3)(\sqrt{y} - 1) + (2/3)(\sqrt{z} - 1) \geq (2/3)(\sqrt{y} - 0) + (1/3)(\sqrt{z} - 0)$$

This expression is called an *incentive compatibility constraint* in the formal theory of incentives. It represents a constraint on the design of a compensation scheme when the principals want to elicit a high level of effort. Note the parallel with the incentive compatibility constraints that arise in problems with precontractual private information (see Chapter 5). As there, the constraint relates the utility of the agent when he or she acts in the desired fashion to that when he or she misbehaves and it thereby limits the arrangements that will work.

With a bit of algebra, the incentive constraint is transformed into the following simpler form:

$$(1/3)\sqrt{z} - 1 \geq (1/3)\sqrt{y} \tag{6.1}$$

According to Equation 6.1, the pay for the good outcome must sufficiently exceed that for the bad in order to compensate the agent for providing the extra effort that makes the good outcome more likely.

Another kind of constraint on the design of compensation is called the *participation constraint*. The terms of employment, taken together, must provide the agent with at least as much utility as that available in outside opportunities. Otherwise, the agent will not agree to *participate* by accepting employment. Recall that for this example, we assume the expected utility of outside opportunities is 1. Then, the participation constraint is:

$$(1/3)(\sqrt{y} - 1) + (2/3)(\sqrt{z} - 1) \geq 1 \tag{6.2}$$

The principal's problem is to find the values of $y$ and $z$ that satisfy these constraints and give the maximum expected returns net of pay to the agent. Furthermore, both $y$ and $z$ must be positive: The agent cannot (in this example) be made to pay the principal when things go badly.

The two constraints are graphed in Figure 6.1. The values of $y$ and $z$ that meet the incentive constraint (6.1) are those above the upward-sloping line, whereas those that satisfy the participation constraint (6.2) are above the downward-sloping line. The shaded area represents the values of $y$ and $z$—the pay levels for bad and good outcomes—that both attract the agent to take the job and motivate him or her to perform as desired. Because the principal's net return is largest when the expected pay to the agent is smallest, the best contract from the principal's point of view is $y = 0$, $z = 9$. This meets both constraints and gives the principal an expected return of $(1/3)(10 - 0) + (2/3)(30 - 9) = (52/3)$.

The agent is no better off than he or she would be if $e$ were observable, whereas the principal's payoff has fallen from (58/3) to (52/3) because the expected wage paid the agent has risen from 4 to 6. The additional expected wage merely serves to compensate for the risk the agent faces. Thus, the unobservability of effort and the consequent moral hazard has an efficiency cost. In this case, the cost arises from having to load too much risk on the agent.

We also need to make sure that it is actually worthwhile to provide incentives to get the agent to select the high effort level. If instead the principal decided to settle for $e = 1$, then there would be no need to use incentive payments. Paying the agent a constant wage of 1—independent of the outcome—would provide the agent with just enough expected utility to get him or her to agree to the contract. Because it would put no risk on the agent, this pay scheme minimizes the costs of employing him or her. Of course, with the agent's pay unaffected by the outcome, the agent will minimize his or her effort provision and pick $e = 1$, as planned. This yields a payoff to the agent of 1 again, and to the principal of $(2/3)(10 - 1) + (1/3)(30 - 1) = (47/3)$. Because this is less than the payoff when $e = 2$ is induced, motivating the higher effort level is worthwhile.

This, of course, is just one example. In other examples, trying to motivate a high level of effort is inefficient because the costs of loading enough risk on the agent to provide the requisite incentives exceed the gains from the higher level of effort. Moreover, this can happen even when a high level of effort would be optimal with full observability and no moral hazard. In the present example, taking the receipts in the good event to be 20 rather than 30 yields such a case. It is also easy to concoct other examples with several possible effort levels where the solution is to give up on trying to induce the level of effort that would be optimal without the observability difficulties, though it is still worthwhile to provide incentives for some effort beyond the minimum conceivable. In these cases, the inefficiency manifests itself both in the agent's bearing risk that he or she would rather avoid and that the principal is better equipped to face, and in the level of effort induced being less than is desirable.

Designing real incentive contracts involves much more complex issues than this example can reveal. Some of these are developed in the next chapter, and more are explored in Chapters 8, 12, and 13.

Part

# IV

# EFFICIENT INCENTIVES: CONTRACTS AND OWNERSHIP

# 7

# RISK SHARING AND INCENTIVE CONTRACTS

*Well, then, says I, what's the use you learning to do right when it's troublesome to do right and ain't no trouble to do wrong, and the wages is just the same?*

*Huckleberry Finn*[1]

In Chapter 6, we examined how insurance of various forms can combine with difficulties of monitoring actions or verifying information to blunt individual incentives. We also surveyed a number of responses to such moral hazard problems. Among these were *incentive contracts*, under which individual incentives are strengthened by holding people at least partially responsible for the results of their actions, even though doing so exposes them to risks that could be more easily borne by an insurance company. In this chapter, we develop a detailed theory of the nature and form of efficient incentive contracts in the presence of moral hazard, establishing a number of general principles that can be used to understand, evaluate, and design such contracts. Although we develop this theory largely in terms of employment contracting and performance pay, the principles are broadly applicable to a wide variety of institutional contexts.

## INCENTIVE CONTRACTS AS A RESPONSE TO MORAL HAZARD

In both theory and practice, there are more options open to society than to insure a risk fully or not to insure it at all. Actual insurance contracts are also incentive contracts: They have provisions that restrict and condition claim payments in ways that provide better incentives than full insurance without removing the essential part of the insurance coverage. The deductible clause that is common in homeowners' fire and theft insurance policies requires the policyholders to bear the initial part of any loss they may incur while still protecting them against large financial losses. Health-insurance policies often require copayments, according to which the insurance pays only a fraction of the costs, with the rest being borne by the insured. Automobile insurance is experience rated, so that those who are responsible for traffic accidents pay higher rates. These features are designed to encourage the insureds to take care and to deter their excessive use of the insurance. For example, the copayments on emergency room visits are set so that, rather than automatically rushing to the emergency room, a patient will wait to be treated in the doctor's office for illnesses and injuries that are not extremely urgent. In this spirit, the policy may provide no coverage at all for treatments that are considered to be elective, such as cosmetic surgery other than that which is necessary to repair damage caused by an injury. As these examples make clear, insurance contracts are designed with profound attention for the need to reduce the waste caused by moral hazard.

Similar moral hazard issues must be faced when devising compensation contracts for employees in a firm. Here, too, there is a balance that needs to be struck between providing incentives and insulating people from risk. To provide incentives, it is desirable to hold employees *responsible* for their performance; this means that employees' compensation or future promotions should depend on how well they perform their assigned tasks. As we will see, however, holding employees responsible typically will involve subjecting them to risk in their current or future incomes. Because most people dislike bearing such risks and are often less well equipped to do so than are their employers, there is a cost in providing incentives. *Efficient contracts balance the costs of risk bearing against the incentive gains that result.*

### Sources of Randomness

If employees were always able to perform as required and if it were easy to determine precisely whether they have behaved as they were supposed to, having pay depend on performance would not generate any risk-bearing costs. An employee could choose whether to perform appropriately or not. Appropriate behavior would be compensated as agreed; inappropriate behavior would go uncompensated and might be penalized. Higher levels of required performance would be associated with higher pay to compensate for the additional effort that the employee is called upon to expend, but there would be no risk in the employee's pay because the outcome is completely under the employee's control.

In most real situations, however, attempts to impose responsibility on employees for their performance do expose them to risk because perfect measures of behavior are hardly ever available. For example, if the employee is expected to give expert advice on some matter, it may be impossible to determine whether the advice is based on the best available information and analysis and whether the recommendations are actually designed to promote the employer's interests, or whether the employee has acted selfishly or deceptively. When care and effort are wanted, it may equally be impossible to determine if employees are doing what they should or slacking off. In these kinds of situations, even though the quality of effort or the accuracy of information cannot itself be observed, something about it can frequently be inferred from observed results, and compensation based on results can be an effective way to provide incentives. Piece rates are a prime example: Rather than trying to monitor directly the effort that the employee provides, the employer simply pays for output.

However, results are frequently affected by things outside the employee's control

[1] The Adventures of Huckleberry Finn, Mark Twain (1884).

that have nothing to do with how intelligently, honestly, and diligently the employee has worked. Sales at a fast-food restaurant may be lower than expected due to the outlet manager's lack of creativity in devising promotional efforts or negligence in supervising the staff, but the low level may also be caused by other factors. Road construction could have made the location less accessible to customers. The opening of a competing restaurant nearby could be to blame. Population growth may have been less than forecast. In the case of a franchise, the franchisor's failure to provide attractive menus or timely deliveries of food could be responsible. Or some combination of these and other factors might be at work. Similarly, if an aircraft crashes, pilot error may be to blame, or poor maintenance, or a design flaw in the craft itself, or a bolt of lightning, or an air traffic control error, and so on. When rewards are based on results, uncontrollable randomness in outcomes induces randomness in the employees' incomes.

A second source of randomness arises when the performance itself (rather than the result) is measured, but the performance evaluation measures include random or subjective elements. For example, the way an employee is evaluated may depend on his or her supervisor's subjective perception of the employee's attitude towards the job and behavior towards other workers. Employees may see this sort of evaluation as a source of risk because it is based partly on elements outside the employee's control. A worker's performance may be evaluated by sporadic monitoring, and these random observations may not give a perfect reflection of the actual quality of the work. In either case, the imperfect evaluation of performance induces randomness in rewards.

A third source of randomness comes from the possibility that outside events beyond the control of the employee may affect his or her ability to perform as contracted. Health problems may reduce the employee's strength and ability to work, concerns about family finances may make it impossible to concentrate effectively on the tasks at hand, or weather or traffic conditions may render meeting a regular schedule impossible. Thus, performance itself becomes random, and so too does performance-based compensation. Consequently, making employees responsible for performance subjects them to risk.

### Balancing Risks and Incentives

It might be possible to insulate employees from these risks by making their compensation absolutely risk free and unrelated to performance or outcomes. In that case, however, the employees would have little direct incentive to perform in more than the most perfunctory fashion, because there are no rewards for good behavior or punishments for bad. As we will see, both here and in Chapter 12 (where we examine compensation issues more specifically) effective contracts balance the gains from providing incentives against the costs of forcing employees to bear risk.

The same considerations arise in many other business transactions. The size of the crop produced by a sharecropper is influenced by weather and pests as well as by the sharecropper's own skill and effort. Traditionally, landowners make part of the sharecropper's compensation proportional to the size of the crop. This arrangement provides helpful incentives that induce the sharecropper to plant drought- and pest-resistant varieties, to irrigate and care for the crops, and so on. However, it also exposes the sharecropper to the risks of a poor harvest—a risk that is at least partially outside his or her control. Similarly, in the United States, a lawyer who sues for damages on behalf of a client often receives a contingency fee (a percentage of the damage award or settlement). This system provides the litigator with an incentive to work hard on behalf of the client, but because the outcome of the lawsuit is not entirely under the litigator's control, both the litigator's income and the client's are uncertain.

Although all of these cases share certain common features, the accuracy of the performance assessments that can be achieved and the need for and possibility of risk sharing or insurance vary from case to case. Because of these differences, the institutions and practices that best balance risk and incentives also vary.

The conclusion that arrangements should vary from case to case is too vague to be of any use to managers or interest to economists. Fortunately, we can do better. The principles developed in this chapter make it possible to reach a relatively subtle understanding of how *optimal* practices can be designed that trade off the value of protecting people from risk against the need to provide them with incentives.

In order to analyze how rational people respond to incentives in insurance-like contracts, we must first examine how rational people behave and interact in risky situations. This involves three steps. The first is to describe the risks precisely, using the language of statistics. Then, we describe how rational people, acting individually, can choose consistently among risky choices and how varying individual attitudes toward risk taking can be incorporated into the analysis. Finally, we examine how groups of people can share risks and form insurance pools, being careful to quantify the benefits of insurance coverage. Given this background, we then examine how people respond to incentives in risky situations. This then allows us to develop the principles of efficiently designed incentive contracts.

## Decisions Under Uncertainty and the Evaluation of Financial Risks

The first element we need is a theory of decisions under uncertainty. There are, in fact, a number of rich theories addressing this subject in great generality, but for our purposes it is enough to consider the special case in which the risks are financial. The first step is to describe the financial risk. We do this using two ideas familiar from statistical theory: the concepts of **mean** and **variance**. These terms are defined in the appendix to the chapter. Here, we illustrate their meaning by computing the mean and variance in an example.

### Computing Means and Variances

Recall that the mean or expected value of a random income is simply the expected amount of income, computed as the weighted average of the possible values that income might take on, with the weight on each value being the probability of that value occurring. The relevant calculations are illustrated in Table 7.1.

The table shows a hypothetical situation in which there is an investment for which the returns are zero with probability one half, \$3,000 with probability one third and \$6,000 with probability one sixth. The *mean* or *expected value* of the return is $\frac{1}{2}(\$0) + \frac{1}{3}(\$3{,}000) + \frac{1}{6}(\$6{,}000) = \$0 + \$1{,}000 + \$1{,}000 = \$2{,}000$. In the table, the calculation works by multiplying the entries in columns 1 and 2 to obtain column 3, and then summing the column. Having higher probabilities on higher values increases the mean.

The *variance* of income is a measure of its variability or randomness. It is computed in columns 4 and 5 of the table. In column 4, we take each possible value, subtract the mean (to get a measure of how far the particular value deviates from the expected value), and square the result (so that terms greater than the average that result from higher-than-expected incomes do not cancel out the negative terms that result when income is less than expected). In column 5, these squared variations are multiplied by the corresponding probability. Summing the column gives the variance. In the example, the variance is $\frac{1}{2}(0 - 2{,}000)^2 + \frac{1}{3}(3{,}000 - 2{,}000)^2 + \frac{1}{6}(6{,}000 - 2{,}000)^2 = \frac{1}{2}(4{,}000{,}000) + \frac{1}{3}(1{,}000{,}000) + \frac{1}{6}(16{,}000{,}000) = 5{,}000{,}000$. (The units are "dollars squared.") If income is certain, then the variance is zero, because the income never deviates from its expected value. Increasing the probability of very high and very low values tends to increase the variance.

Table 7.1 **Sample Computation of Mean and Variance**

| 1<br>Probability | 2<br>Return | 3<br>(1) × (2) | 4<br>(Return − Mean)² | 5<br>(1) × (4) |
|---|---|---|---|---|
| 1/2 | 0 | 0 | 4,000,000 | 2,000,000 |
| 1/3 | 3,000 | 1,000 | 1,000,000 | 333,333 |
| 1/6 | 6,000 | 1,000 | 16,000,000 | 2,666,667 |
| | | Mean = 2,000 | | Variance = 5,000,000 |

## Certainty Equivalents and Risk Premia

One of the main hypotheses we employ in this chapter is that most people are *risk averse*; that is, they would prefer receiving a certain income of $\bar{I}$ to receiving a random income with expected value $\bar{I}$. The amount the person would be willing to pay to make the switch is the **risk premium** associated with the random income. The magnitude of the risk premium depends on both the riskiness of the income and the individual person's degree of risk aversion. The amount that is left after the risk premium is paid is the **certainty equivalent** of the random income. It is the amount of income, payable for certain, that the person regards as equivalent in value to the original, random income.

One of the central results of decision theory is that the certainty equivalent can be estimated by a simple formula: $\bar{I} - \frac{1}{2}r(\bar{I})Var(I)$, where $\bar{I}$ and $Var(I)$ are the mean and variance of the random variable $I$, and $r(\bar{I})$ is a parameter of the decision maker's personal preferences called the **coefficient of absolute risk aversion** for gambles with mean $\bar{I}$. The mean in this formula is the mean income, and the amount subtracted from it in the formula is the risk premium; it is equal to one-half times the coefficient of absolute risk aversion times the variance of the income. According to the formula, the risk premium is proportional to the coefficient of absolute risk aversion: People who are more risk averse according to this measure are willing to pay proportionately larger risk premiums to avoid a given risk. If the coefficient of absolute risk aversion is zero, then the person is unwilling to pay any premium to avoid the risk. Such a person is called **risk neutral**. A person is risk averse when the coefficient of absolute risk aversion is positive. The amount $\bar{I} - \frac{1}{2}r(\bar{I})Var(I)$ that is left in expectation after the risk premium is deducted is called the person's certain equivalent income or the certainty equivalent of the random income $I$.

If there is no uncertainty regarding the level of income, then $Var(I) = 0$; the only value that income actually might take on is $I = \bar{I}$. Then, the formula yields the sensible result that the person is as well off with the nonrandom income $I$ as with a certain amount that is equal to $I$: The thing is as good as itself. When $I$ does vary (so $Var(I)$ is positive) and the person is risk averse (so $r(\bar{I})$ is also positive), the risk premium is positive. This means that he or she would be willing to accept a lower amount than $\bar{I}$ to avoid the risk. More precisely, the risk premium, $\frac{1}{2}r(\bar{I})Var(I)$, is the amount that the person would pay to have the certain income $\bar{I}$ for sure rather than face the uncertainty in $I$.[2]

[2] The estimate of the certainty equivalent given in this formula is good when the variance is not too large or the coefficient of risk aversion is small. In terms of the example in Table 7.1, where the mean income was $2,000 and the variance was 5,000,000, the formula becomes $2,000 - 2,500,000r(\bar{I})$. This approximation is reasonable only for values of $r$ in the range of .00008 or less (corresponding to a risk premium of 200); when $r \leq .0008$, it yields the nonsensical answer that the individual would be indifferent between the gamble and getting a negative income for sure. In using the approximation, we thus assume that the variance of the uncertain income is not too large relative to the individual's risk aversion.

## Risk Premia and Value Maximization

Our analysis in this chapter uses the value maximization principle, which in the context of uncertainty asserts that an arrangement is efficient if and only if it maximizes the total certain equivalent wealth of all the parties involved. Recall from Chapter 2 that the premises needed to derive the principle are (1) that each person has enough wealth to make whatever payments might be called for under any relevant contract and (2) that each person has a well-defined willingness to pay for any given product or service and the amount of this monetary valuation does not depend on his or her income level. As discussed in Chapter 2, these are strong and often unrealistic assumptions, but they greatly simplify the analysis and enable us to separate analytically the effects of the level and variability of income from all other effects on the matters of interest. In the context of uncertain income, the second assumption is reduced to this: The risk premium that a person would pay to eliminate a given amount of variance must not depend on the expected level of income $\bar{I}$. In view of the risk premium formula, this means that $r(\bar{I})$ must not depend on $\bar{I}$. Throughout the rest of this chapter, we make that assumption and write $r$ instead of $r(\bar{I})$. With this assumption, the crucial formulas become:

$$\begin{aligned} \text{Expected Income} &= \bar{I} \\ \text{Risk Premium} &= \tfrac{1}{2}rVar(I) \\ \text{Certain Equivalent} &= \bar{I} - \tfrac{1}{2}rVar(I) \end{aligned}$$

We use these formulas to calculate the benefits of insurance and the costs of the risk bearing that is required to provide incentives.

# Risk Sharing and Insurance

One of the most fundamental facts about the economics of risk is that when several people are facing statistically independent risks, then by sharing the risks among themselves they can greatly reduce the cost of risk bearing. Two risks are **statistically independent** if knowing the realized value of one risk gives you no information about the value that the other will achieve. For example, the amount you won or lost per dollar invested in the state lottery today does not give you any reason to change your estimates of the likely returns in the stock market tomorrow. In contrast, for risks that are not independent, knowledge of one *is* useful in predicting the other. For example, the prices of gold on the London and New York markets are both random, but they tend to move together under the influence of arbitrage (buying in one market and selling in another to make a riskless profit). Thus, knowing the price in London tells you something useful about what the New York price is likely to be, and so the two risks are not independent. This **principle of risk sharing**—that sharing independent risks reduces the aggregate cost of bearing them—is the basis of all financial insurance contracts.

## How Insurance Reduces the Cost of Bearing Risk

In modern economies there are many kinds of institutions to assist people in sharing risks. One important group consists of the insurance companies. Having many policyholders, the insurance companies can spread risks very widely, enabling the companies to reduce individual risks greatly. If the risks are independent and the number of policyholders quite large, the risks are effectively eliminated and insurance works very well. For example, the risk that you will suffer an automobile accident is

very nearly independent of the risk that any other particular person will do so, therefore automobile insurance is a feasible enterprise. Insurance companies specialize in evaluating individual risks and, by pooling the risk-bearing capacity of policyholders and (sometimes) shareholders, they reduce the cost of the risk bearing to negligible proportions. Pooling independent risks also has the additional advantage of making the insured losses statistically predictable. An insurance company can ask each insurance policyholder to pay a price for insurance equal to the expected amount of the loss, plus a margin for expenses and profit, and can be reasonably sure that the aggregate premium income together with a proportionately small reserve fund will enable it to pay for whatever losses may be suffered, even in a bad year.

Some kinds of risks, however, are so large and pervasive in their impact that they cannot be made negligible by sharing and they cannot be managed by traditional insurance arrangements. (Technically, the risks that people bear in this case are not statistically independent.) For example, an oil price increase would have such widespread effects, reducing the effective incomes of most people in oil-consuming countries, that no amount of risk sharing among those oil consumers can insulate them from the loss. Risks of this general kind are shared through other markets, especially the financial markets. By purchasing stock in companies that own oil reserves, for example, an investor who is especially vulnerable to oil price increases can arrange to have an offsetting profit if oil prices increase. Financial markets allocate many other kinds of risks, as well. For our purposes, an important example is the investment risks that are taken by firms, such as those associated with a new technology. The risk of failure of the technology is borne by shareholders in the company that develops it, and this capacity for risk sharing reduces the firm's cost of financing the investment, helping to promote technical change.

## Efficient Risk Sharing: A Mathematical Example

Suppose that there are two people, A and B, each of whom has some risk associated with his or her income, where these risks are independent. Let $I_A$ and $I_B$ represent their random incomes, with means $\bar{I}_A$ and $\bar{I}_B$ and variances $Var(I_A)$ and $Var(I_B)$, and let $r_A$ and $r_B$ denote their coefficients of absolute risk aversion. In view of our earlier assumption, the value maximization principle applies. Consequently, every efficient risk-sharing contract maximizes the total certain equivalent income of all the parties, and every such contract is an efficient one.

If the parties make no special arrangements, then the total cost they suffer on account of risk bearing, that is, the total risk premium, is $\frac{1}{2}r_A Var(I_A) + \frac{1}{2}r_B Var(I_B)$, which is the sum of the two individual risk premia. Suppose that the parties instead agree on a risk-sharing contract with party A receiving a fraction $\alpha$ of the income $I_A$ and ß of the income $I_B$ (and thus of the risks associated with the two uncertain incomes.) In addition, suppose A receives a cash transfer of $\gamma$ for the risk-sharing services provided. (This transfer might be positive or negative, but it is independent of the actual, realized incomes.) Party B receives the remaining share of each risk and makes the cash payment $\gamma$. After this agreement, A's income will be $\alpha I_A + ßI_B + \gamma$ and B's will be $(1 - \alpha)I_A + (1 - ß)I_B - \gamma$. This is a feasible agreement because the total income each party receives always adds up to $I_A + I_B$, the amount available. With this agreement, the *total risk premium* of the two parties is:

$$\text{Total Risk Premium} = \tfrac{1}{2}r_A Var(\alpha I_A + ßI_B + \gamma) + \tfrac{1}{2}r_B Var((1 - \alpha)I_A + (1 - ß)I_B - \gamma) \quad (7.1)$$

Because the total certain equivalent income of the two parties is equal to the mean income, $\bar{I}_A + \bar{I}_B$, minus the risk premium, the efficient arrangements are those that minimize Equation 7.1.

Using identities about variances (see Formula 7.18 in the appendix), Equation 7.1 is a quadratic function of $\alpha$ and ß. The total risk premium is minimized when $\alpha/(1 - \alpha) = ß/(1 - ß) = r_B/r_A$. For example, suppose $r_A = 2$ and $r_B = 4$. The higher value for B's coefficient of absolute risk aversion indicates that B finds bearing risk more onerous than does A. Indeed, the risk premium that B attaches to any given risk is twice the amount A would pay to avoid the risk. In these circumstances, we might expect that A would bear more of the risk than would B. Evaluating the solution, we see that $\alpha/(1 - \alpha) = ß/(1 - ß) = 2$, so $\alpha = ß = \frac{2}{3}$ and $(1 - \alpha) = (1 - ß) = \frac{1}{3}$: A does in fact bear most of both risks. Moreover, A bears the same share (two thirds) of both.

To formulate the general principle that applies here, it is helpful to think in terms of different peoples' capacity to bear risk. We measure this by introducing the notion of risk tolerance. Someone with a coefficient of absolute risk aversion of $r$ will be said to have **risk tolerance** of $1/r$. Notice that in the preceding example, A's share of each risk is equal to A's share of the total risk tolerance ($\frac{2}{3} = \frac{1}{2}/(\frac{1}{2} + \frac{1}{4})$).

These calculations actually reflect a general principle that can be shown to hold for any number of people and any number of financial risks: *When risks are shared efficiently, the share that a party bears in each risk is the same and is equal to his or her share of the total risk tolerance of the group.* Moreover, when risks are allocated efficiently, the total risk premium comes out to be:

$$\text{Total Risk Premium} = \tfrac{1}{2}Var(I_A + I_B)/[(1/r_A) + (1/r_B)] \quad (7.2)$$

Equation 7.2 resembles the formula for the risk premium charged by a single decision maker. It says that when risks are shared efficiently among a group of people, the total risk premium is the same as if the total risk were borne by a single decision maker whose risk tolerance is the sum of the members' individual risk tolerances. In the preceding numerical example, $(1/r_A)+(1/r_B) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$. This formula, too, is general; it can be shown to hold for any number of people and any number of financial risks. With efficient risk sharing, the group is less risk averse than the people comprising it and so the costs of bearing risks can be reduced.

When individual risks are independent, these facts imply that sharing risks can be a very effective way to reduce the cost of risk bearing. For example, if there are $n$ people, each with an income with variance $v$ and each with the same coefficient of risk aversion $r$, and if each bears the income risk separately, then the risk premium will be $\frac{1}{2}rv$ per person. If the people share the income risks efficiently, then each will have a $1/n$ share of the total risk. The variance of the total risk is $V = nv$, so the variance of an individual $1/n$ share is $V/n^2 = v/n$ (see Formula 7.18 in the appendix again). Therefore, by sharing risks, each person's risk premium is reduced from $\frac{1}{2}rv$ to $\frac{1}{2}rv/n$. When $n$ is large, even substantial financial losses can be reduced to economic insignificance by sharing them efficiently across the group.

## Optimal Risk Sharing Ignoring Incentives

For both insurance companies, with their wide base of policyholders, and publicly traded corporations, with their wide base of shareholders, it is reasonable to suppose as a first approximation that the total risk tolerance of the company is infinitely larger than the risk tolerance of any individual policyholder or employee. As we mentioned earlier, an institution or person with infinite risk tolerance is said to be risk neutral: The coefficient of absolute risk aversion is zero and so the risk premium for bearing any risk is also zero. Applying our general propositions to the case where risks are to be shared between a risk-neutral insurance company and a risk-averse insurance policyholder or between a large, risk-neutral employer and a risk-averse employee, we find that the optimal share of the risk to be borne by the insurance buyer or employee

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:

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

Compensation thus consists of a base amount, $\alpha$, plus a portion that varies with the observed elements, $z$ and $y$. We use ß 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 ß. 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 ß dollars, so higher levels of ß bring greater returns to increased effort.

The parameter $\gamma$ indicates how much relative weight is given to the information variable $y$ (as compared to $z = e + x$) in determining compensation. If $\gamma$ is set at zero, then $y$ is not used in determining compensation. Given any value for $\gamma$, 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 $\gamma$.

As an example of such a contract, suppose $\alpha$ is \$10,000, ß is \$20 and $\gamma$ is 0.5. Then expected pay is \$10,000 + \$20$e$, 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 \$13,000. 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 ß 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 ($e$, $\alpha$, ß , $\gamma$) 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: $\alpha + ß(e + \bar{x} + \bar{y}) - C(e) - \frac{1}{2}rVar[\alpha + ß(e + x + \gamma y)]$, where $\bar{x}$ and $\bar{y}$ 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 $\bar{x}$ and $\bar{y}$ 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:

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

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

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

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

Notice that the employee's certain equivalent consists of $\alpha$ plus a function of the other variables (ß, $\gamma$, $e$) 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 ($\alpha$, $\beta$, $\gamma$) 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 $\beta$; 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 ($e$, $\alpha$, $\beta$, $\gamma$) 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 $\alpha$, $\beta$, and $\gamma$ 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 $\beta$ 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 $\beta$ 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 $\alpha$ 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 $\alpha$ 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 $\alpha$. As for $\gamma$, it is clear that the total certain equivalent is maximized when $\gamma$ 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.

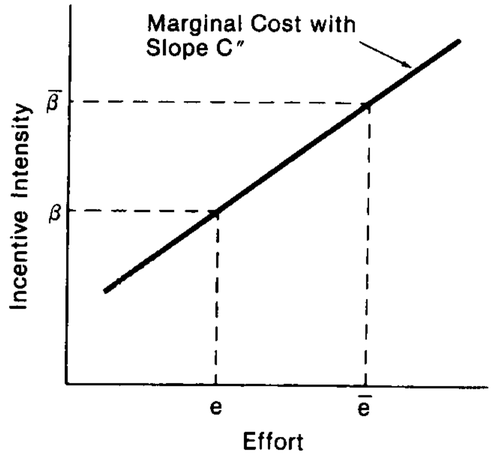


**Figure 7.1:** Increasing effort provided from $e$ to $\bar{e}$ requires increasing $\beta$ to $\bar{\beta}$, where $\bar{\beta} - \beta = (\bar{e} - e)C''$.

## The Informativeness Principle

This last result—that $\gamma$ 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 $\gamma$ that makes $Var(x + \gamma y)$ smaller than $Var(x)$, the estimate that results when $y$ is ignored and $\gamma$ is set at zero. The optimal value for $\gamma$ 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 $\gamma$ yields the result that $\gamma$ 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, $\gamma$ 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 $\gamma$ 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 $\gamma$ 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 $\gamma$ 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 $\gamma$.

Also note that as the variance of $y$ increases, the magnitude of $\gamma$ optimally decreases. Larger values of $Var(y)$ mean more "noise"—less reliable information—and the optimal choice of $\gamma$ 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 $\gamma$ 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: $ß = P'(e)/[1 + rVC''(e)]$.

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

the results are profitable. For example, it is counterproductive to use incentives to encourage production workers to work faster when they are already producing so much that the next stage on the production line cannot use their output. According to the incentive intensity principle, the optimal intensity is proportional to the profitability of incremental effort, provided the other three factors remain unchanged.

The second factor is the risk aversion of the agent. The less risk averse the agent, the lower the cost he or she incurs from bearing the risks that attend intense incentives. According to the incentive intensity principle, more risk averse agents ought to be provided with less intense incentives.

The third factor is the precision with which performance is measured. Low precision corresponds to high values of the variance V, which according to the formula means that only weak incentives should be used. It is futile to use wage incentives when performance measurement is highly imprecise, but strong incentives are likely to be optimal when good performance is easy to identify.

The final factor is the responsiveness of effort to incentives, which is inversely proportional to $C''(e)$ (see Figure 7.1). For example, an employee working on a fixed rate production line cannot increase his or her own output in response to piece rate incentives. According to the incentive intensity principle, incentives should be most intense when agents are able most able to respond to them. Generally, this happens when they have discretion about more aspects of their work, including the pace of work, the tools and methods they use, and so on. An employee with wide discretion facing strong wage incentives may find innovative ways to increase his or her performance, resulting in significant increases in profits.

### *Mathematical Derivation of the Optimal Incentive Intensity*

Figure 7.2 illustrates the trade-offs that determine the optimal intensity. The intensity, ß, is measured on the horizontal axis and its marginal benefits and costs on the vertical axis. The downward-sloping line records the net marginal benefit of increasing the intensity of incentives. The net marginal benefit of extra effort is $P'(e) - C'(e)$. To determine the net marginal benefit of incentives, the marginal benefit of effort must be multiplied by the rate at which extra effort is supplied for each extra unit of intensity. That rate, as we have previously seen, is $1/C''(e)$. Since the agent will choose $e$ so that $ß = C'(e)$, the net marginal benefit is $(P'(e) - C'(e))/C''(e) = (P'(e) - ß)/C''(e)$, as shown in the Figure. The transaction cost associated with setting effort intensity ß is the risk premium $\frac{1}{2}rVß^2$, with associated marginal cost $rVß$, as plotted in the Figure. The optimal intensity of incentives occurs at the point where the marginal benefit and marginal cost are equal.

To find the optimal intensity by direct maximization, write the total certain equivalent for any fixed value of $e$ and ß as $P(e) - C(e) - \frac{1}{2}rß^2V$, by Equation 7.4b. From the incentive constraint of Equation 7.5, we know that $ß = C'(e)$, so the objective can be rewritten as:

$$\text{Total Certain Equivalent} = P(e) - C(e) - \tfrac{1}{2}rC'(e)^2V \qquad (7.6)$$

Equation 7.6 gives a clear picture of the benefits enjoyed and costs incurred for any given level of effort. The benefit term in this equation is just the profit $P(e)$, but the cost has two components: the direct cost $C(e)$ incurred by the agent plus the transaction cost $\frac{1}{2}rC'(e)^2V$ of providing the requisite incentives.

The optimal level of effort $e$ under the contract is found by differentiating the total certain equivalent with respect to $e$ and setting that derivative equal to zero: $0 = P'(e) - C'(e) - rVC'(e)C''(e)$. Using Equation 7.5 again, we can replace $C'(e)$ by ß in this expression to obtain: $0 = P'(e) - ß - rVßC''(e)$. Solving this for ß results in the formula given in the incentive intensity principle.

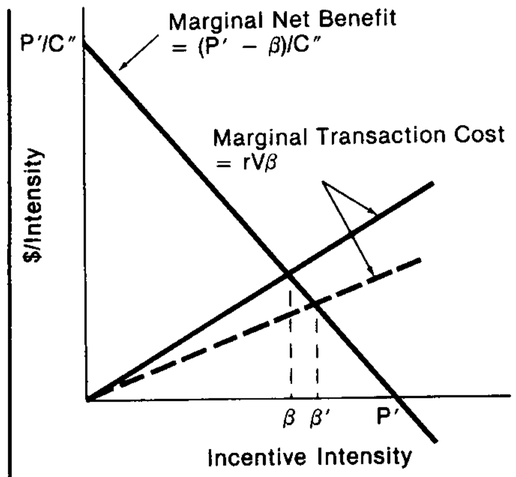


**Figure 7.2:** The optimal intensity of incentives balances the direct net marginal benefits of increasing ß against the marginal transaction cost.

APPLICATION: INCENTIVES FOR JAPANESE SUBCONTRACTORS Two recent studies have been performed that compare the recommendations of the incentive intensity principle with the actual contractual practices used to compensate subcontractors who supply parts or components for large Japanese automobile and electronics firms.[3] In Japanese practice, the amount paid by a manufacturing firm for its inputs depends on the actual costs as measured in the supplier company's accounting records, rather than being a contractually fixed price. If the target level of cost is $\bar{x}$ and the actual cost incurred is $x$, then the supplier is paid $x + ß(\bar{x} - x)$. That is, the manufacturing firm pays the actual cost incurred plus a fraction of the difference between the target cost (which is negotiated to include an allowance for profit) and the realized cost. This adjustment is an incentive term. If the supplier's actual cost is less than the target, it gets to keep some of the savings. If its costs exceed the target level, then the manufacturing company absorbs some of the difference. Thus, if the actual cost $x$ is less than the target , the subcontractor earns an extra profit of $ß(\bar{x} - x)$; if it is more, then $ß(\bar{x} - x)$ is negative, which means that the subcontractor pays a penalty for its poor performance.

To analyze this case, notice that an effort that reduces costs by 1 yen also adds 1 yen to the manufacturing firm's profit, so we may take $P'(e) = 1$. Consequently, the theory recommends that $ß = 1/(1 + rVC'')$. The researchers rearranged the terms in this equation to obtain $1/ß - 1 = rVC''$. Taking logarithms of both sides of the new equation leads to an equation that the researchers could test using linear regression analysis:

$$log(1/ß - 1) = log(r) + log(V) + log(C'') \qquad (7.7)$$

The ideal would now be to use data on ß, $r$, V and $C''$ from different contracts to estimate the empirical relationship among these variables. Then one could test statistically whether the empirical relationship was the one predicted by the theory.

---

[3] S. Kawasaki and J. McMillan, "The Design of Contracts: Evidence from Japanese Subcontracting," *Journal of Japanese and International Economies*, 1 (1987), 1327–49; and B. Asanuma and T. Kikutani, "Risk Absorption in Japanese Subcontracting: A Microeconometric Study on the Automobile Industry," forthcoming in the *Journal of Japanese and International Economies* (1991).

However, the available data did not provide direct information on all these variables.[4] In such a situation, the next best thing is to identify instruments for the theoretical variables of interest, which are *log*(*r*), *log*(V) and *log*(C″). An *instrument* for a variable is another variable that (1) can be observed, (2) varies directly with the actual variable of interest, and (3) is uncorrelated with the other variables of interest.

To test Equation 7.7, the researchers first estimated $1 - ß$ by dividing the variation in the supplier's profits over time by the variation in their costs. These estimates were then used to tabulate $log(1/ß - 1)$ for the various firms in the sample. The risk aversion *r* was assumed to be inversely proportional to various measures of the size of the firm, such as the number of the firm's employees. Size variables therefore were used as instruments for *log*(*r*) in the equation. The variance V in costs was estimated by determining the trend in costs and then computing the variation in actual costs around the trend over time. In theory, C″ should be inversely proportional to the scope for performance improvement by the agent. The researchers supposed that the scope was proportional to the firm's value added in the production process (in the Kawasaki and McMillan analysis) or to the firm's responsibility under the contract for supplying technology and designing parts and production processes (in the Asanuma and Kikutani analysis). These value-added and responsibility measures were used as instruments for C″ in the actual estimation. With only these instruments for the actual variables of interest, all that could be hoped for is that the signs of the coefficients in the estimated equations would be the same as predicted by the theory: The intensity of incentives ß should be greater for firms with more employees, more value added, and less variability in year-to-year performance. The empirical findings were consistent with these predictions.

The tests we have described represent only weak evidence in support of the theory. The equation whose coefficients were finally estimated was not the exact one predicted by the theory, and the instruments used are not beyond criticism. Moreover, the estimation procedure did not test whether there were other variables affecting actual choices of ß that were not predicted by the theory and, if so, how important those other variables were for understanding incentives. Nevertheless, the evidence obtained is consistent with the theory: Incentive contracts for Japanese suppliers do appear to depend on the considerations identified by the theory in the general way that the theory predicts.

APPLICATION: INCENTIVES IN OIL AND GAS TAX SHELTER PROGRAMS Another study has tested the incentive-intensity principle in the context of the organization of oil and gas tax shelters in the United States in the early 1980s.[5] At that time, many drilling operations were financed by limited partnerships. As you recall from Chapter 6, under the federal tax laws that then prevailed, the partners could often save on taxes if the limited partners paid all the costs of exploring for oil (which were tax deductible when the costs were incurred), whereas the general partner(s) paid the costs of completing wells in which oil was found (which were "capitalized costs" for tax reporting purposes). The general partner and the limited partners would then share any revenues enjoyed when oil was pumped from producing wells.

A problem with this tax-reduction scheme is that it created a difference in interests between the general partner, who controlled the partnership's activities, and the limited partners because each bore a different kind of expense. If a well were found to have oil, the general partner had to bear 100 percent of the cost of completing the well, but typically received only 25 percent of the oil revenues. Suppose that after the exploration costs have been sunk, a well were found to have only enough oil that the general partner would need to have a 50 percent share of revenues to recover the well-completion costs. Then, it would not be in his or her interest to complete the well, even though the full revenues would more than cover the completion costs.

Several of the prospectuses used by the general partners to attract investors described the problem quite candidly. According to one:

> A situation may arise in which the completion of an initial well (the majority of the costs of which are capitalized costs) on a prospect would be more advantageous to the limited partners than to the general partners. The situation would arise where a completion attempt on an initial well, the majority of the costs of which are paid by the general partners, could apparently result in a marginal well which would return some but not all of the completion cost incurred by the general partners but would return revenue to the limited partners.[6]

The conflict of interest described here is likely to be most severe when many of the wells being drilled are "marginal" prospects. If the well that is found is a gusher, then even the 25 percent of revenues accruing to the general partner would make completion of the well highly profitable. The general partner seen as the agent of the limited partners, therefore, is most likely to be responsive to completion incentives—to have his or her behavior positively affected by explicit incentives—when many of the wells to be completed are marginal ones. No explicit incentives for completing wells are necessary when they are very productive, and giving such incentives would not have much effect on the general partner's behavior. Economic theory predicts that the contracts that are actually used should be responsive to this difference in completion incentives.

To test this theory, the researcher divided drilling programs into three types: exploratory programs, developmental programs, and balanced programs. *Exploratory drilling programs* were ones in which wells were drilled in new areas, where the greatest likelihood was that no oil would be found but any wells that were found were unlikely to be marginal. In these programs, the conflict between the general and limited partners' interests in completing wells was likely to be small, and the general partners' completion decision was likely to be little affected by any special contractual incentives. Some 96 percent of the money invested in these exploratory drilling programs in the sample was in contracts that were designed to minimize taxes, with no special allowances to improve the general partners' completion incentives. *Developmental drilling programs* were ones in which all drilling occurred in an area that had been previously explored and where oil was known to be present, but where no more major finds were expected. Many developmental wells turn out to be marginal wells, so we should expect that the general partner would have been quite responsive to incentives to complete these wells. The researcher found that only 23 percent of the money invested in these programs was in contracts that provided no completion incentives. For *balanced drilling programs*, which contained a mix of exploratory and developmental wells, the corresponding figure was 37 percent.

This evidence provides a useful test of one aspect of the incentive-intensity

---

[4] Kawasaki and McMillan used data reported in *MITI's Census of Manufacturers (The Firm Series)* and *Surveys of Industries*. Asanuma and Kikutani limited their attention to Japanese automobile manufacturers, from whom they could obtain somewhat more detailed information.

[5] Mark Wolfson, "Empirical Evidence of Incentive Problems and Their Mitigation in Oil and Gas Tax Shelter Programs," *Principals and Agents: The Structure of Business*, J. Pratt and R. Zeckhauser, eds. (Boston: Harvard Business School Press, 1985), 101–27.

[6] *Prospectus of the Hilliard Fund* (1982), 22, as quoted by Wolfson.

principle. The impact of any given monetary incentive on the agent's behavior varies with circumstances, and the principle predicts that incentives will be more intense and more often incorporated into contracts when the agent's responsiveness to them is high. The evidence in this case generally confirms the prediction of the principal-agent model: Incentives are provided when they are likely to make a difference.

## The Monitoring Intensity Principle

So far, we have assumed that the measurement of performance is outside the scope of the model; that is, the variance V with which efforts are measured has been treated as outside the employer's control (other than through the determination of γ). Often, however, it is possible for an employer to improve measurement by devoting resources to that objective. For example, in a factory, the number of workers per supervisor could be reduced to allow closer monitoring, or more quality-control tests could be made. For service workers, customers could be interviewed to learn whether they were satisfied with the service. In a telephone ordering or service operation, call-counting and timing equipment could be installed or supervisors could listen in on incoming calls to see how well they are handled. All of these things are costly, but all improve the employer's information about how employees are performing.

To investigate how much should be spent on monitoring, suppose that the variance of the performance measure can be controlled at a cost. Let M(V) be the minimum amount that must be spent on monitoring needed to achieve an error variance as low as V. It is generally costly to reduce the error variance, so we suppose that $M$ is a *decreasing* function—settling for a larger V entails lower monitoring costs. We also suppose that the marginal cost of variance reduction is a rising function, that is, $M'(V)$ is increasing. Rewriting Equation 7.4b to include the cost of the resources that are spent on measurement, we have:

$$\text{Total Certain Equivalent} = P(e) - C(e) - \tfrac{1}{2}rV\beta^2 - M(V) \qquad (7.8)$$

The relationship between $e$ and ß is still determined by the incentive constraint Equation 7.5, which is unaffected by the introduction of costly measurement. We may therefore hold $e$ and ß fixed and choose V to maximize the expression in Equation 7.4b. Taking the derivative of Equation 7.8 with respect to V leads to:

$$-\tfrac{1}{2}r\beta^2 - M'(V) = 0 \qquad (7.9)$$

According to this equation, the marginal cost of reducing V, which is $-M'(V)$—a positive number—must be equal to $\frac{1}{2}r\beta^2$ at the efficient solution.

> **The Monitoring Intensity Principle:** Comparing two situations, one with ß set high and another with ß set lower, we find that V is set lower and more resources are spent on measurement when ß is higher: When the plan is to make the agent's pay very sensitive to performance, it will pay to measure that performance carefully.

The determination of V is illustrated in Figure 7.3. The downward sloping curve gives the marginal cost of reducing the variance with which performance is measured. Because the risk premium is $\frac{1}{2}\,r\beta^2V$, the marginal cost of variance changes is depicted in the figure by a solid line at level $\frac{1}{2}r\beta^2$. When the incentive intensity is reduced from ß to $\bar{\beta}$, the chosen level of V increases: Fewer resources are spent on measurement.

There may appear to be some circularity in our several observations. In the incentive-intensity principle, we claim that ß should tend to be set large when V is low. In the last paragraph, we claim that firms should try to reduce V when ß is large.

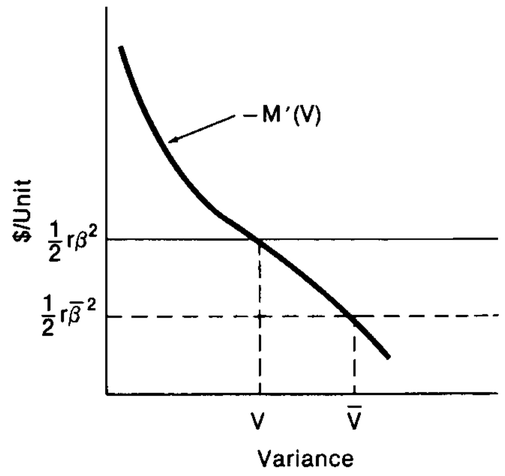


**Figure 7.3:** The optimal level of measurement equates the marginal cost and marginal benefit of variance reduction. Less intense incentives lead to higher V (less measurement).

Which causes which? Do intense incentives lead firms to careful measurement, or does careful measurement provide the justification for intense incentives?

The answer is that, in an optimally designed incentive system, the amount of measurement and the intensity of incentives are chosen together: Neither *causes* the other. However, setting intense incentives and measuring performance carefully are *complementary* activities in the sense described in Chapter 4; undertaking either activity tends to make the other more profitable.

Figure 7.4 illustrates the situation. The two solid lines in the figure depict the two relationships between measurement and incentive intensity just described. One of these lines specifies the optimal intensity of incentives ß for any particular measurement variance; the other specifies the optimal variance for any particular intensity of incentives. Notice that both lines slope downward. According to the incentive-intensity principle, ß falls when the variance V rises. Similarly, according to the monitoring intensity principle, V falls as ß rises; it pays to measure more carefully (lower V) when incentives are intense. The point where the two lines cross determines the optimal combination; it is the point where V is chosen optimally for the given intensity of incentives and ß is selected optimally for the given measurement error.

The dotted line in Figure 7.4 shows how ß would depend on V in different circumstances, in which $P'$ was higher or $C''$ lower. According to the incentive-intensity principle, these changes would lead to higher levels of ß for any fixed level of V. That change is represented in Figure 7.4 by the dotted line lying to the right

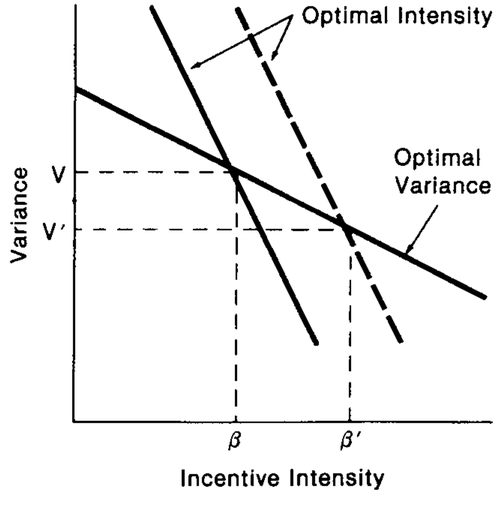


**Figure 7.4:** An increase in $P'$ or a fall in $C''$ leads to more intense incentives and more measurement (less variance).

of the original line determining ß as a function of V. Notice that the point of intersection of the new optimal incentive-intensity line with the optimal variance line is lower and to the right of the original point of intersection. The change leads to sharper incentives and lower variance (more monitoring).

## The Equal Compensation Principle

Now we enrich our conception of behavior in the firm to recognize that most employees do more than one thing as part of their jobs. When there are several activities being conducted, the employer will be concerned that employees allocate their time and efforts correctly among the various things that need to be done. This complicates the problem of providing incentives.

For example, suppose that the marketing representatives for a company making specialty steel alloys perform several kinds of activities: They solicit business from new customers, provide problem-solving services and advice to customers about how to use the company's alloys, gather information about competitors' marketing activities, and report about possible new products that might sell well. Of these several activities, the easiest one to monitor is direct sales efforts, because it leads immediately to measurable sales. Some of the other activities also lead to sales, with some time lag: Keeping customers happy is likely to increase the representative's sales over a period of time. If there is high turnover in the sales jobs, then the information about how well customers are being served may not be available in a timely enough fashion to use for compensating the responsible representative. Finally, some of the activities, such as monitoring competitors' moves, are much more difficult to evaluate than is simple sales performance. If the firm were to compensate the marketing representatives based primarily on the accurately measured current sales figure, that might induce a distortion in their behavior, causing them to switch efforts toward the immediate high-payoff activity of generating sales and away from the activities necessary to keep customers happy and the firm well informed. If that sort of behavior led eventually to a loss of customers and declining sales, the representative could seek another job, proudly displaying the sales performance he or she achieved in the first job. A related problem might arise for the salespeople in a department store, who might be tempted to maximize immediate commissions by pressuring a customer to buy a more expensive product than necessary, leading to dissatisfied customers and lower future sales for the store, not only of that one department's products but also of products sold in other departments.

Alternatively, suppose that a fast-food chain wants its outlets to be profitable but also wants them to contribute to the chain's reputation for cleanliness, fast service, and hot, fresh food, because that reputation enhances sales at *other* outlets. These profit and reputation goals can be in conflict. For example, a fast-food chain outlet along a highway where many of the customers visit only once would suffer little loss of profits if its hamburgers were sometimes cold and its bathrooms dirty, but the chain's other stores might lose business on that account. If the chain compensates the store manager on the basis of sales alone, the manager would be unlikely to take full account of the effects of his or her actions.

These observations lie behind our fourth principle of incentive contracting and compensation.

> **The Equal Compensation Principle**: If an employee's allocation of time or attention between two different activities cannot be monitored by the employer, then either the marginal rate of return to the employee from time or attention spent in each of the two activities must be equal, or the activity with the lower marginal rate of return receives no time or attention.

The equal compensation principle imposes a serious constraint on the incentive-compensation formulas that can be effective in practice. In particular, if an employee is expected to devote time and effort to some activity for which performance cannot be measured at all ($V = \infty$), then incentive pay cannot be effectively used for any other activities that the individual controls. The use of straight salary compensation for managers can often be justified on these grounds.

MATHEMATICS OF THE EQUAL COMPENSATION PRINCIPLE Suppose the employee does two different things, signified by levels of effort $e_1$ and $e_2$. We will think of these levels of effort as time devoted to two activities, and we assume that the cost incurred by the employee is an opportunity cost: It is time that becomes unavailable for other, more pleasant or rewarding activities. It then makes sense to write the cost as depending only on the total effort, not on its division between the two tasks: $C(e_1 + e_2)$. The employer measures performance by observing the indicators $e_1 + x_1$ and $e_2 + x_2$, where $x_1$ and $x_2$ have expected values of $\overline{x}_1$ and $\overline{x}_2$.

Suppose that the employer pays the employee according to a linear compensation formula based on the two indicators: The wage paid is then $w = \alpha + ß_1(e_1 + x_1) + ß_2(e_2 + x_2)$. How should $\alpha$, $ß_1$, $ß_2$, $e_1$, and $e_2$ be chosen?

To take incentives into account in the problem, we first examine the employee's objective given this compensation rule. A self-interested employee will choose $e_1$ and $e_2$ to maximize his or her certain equivalent income:

$$\text{Employee's Certain Equivalent} = \alpha + ß_1(e_1 + \overline{x}_1) + ß_2(e_2 + \overline{x}_2) - C(e_1 + e_2) - \tfrac{1}{2} rVar(ß_1x_1 + ß_2x_2) \qquad (7.10)$$

For this problem, we suppose that the effort is restricted to a nonnegative number: $e_1, e_2 \geq 0$. If $e_1$ is strictly positive, then at the maximizing choice for the employee, the derivative of Equation 7.10 with respect to $e_1$ must be zero, so $ß_1 = C'(e_1 + e_2)$. Similarly, if $e_2$ is strictly positive, then $ß_2 = C'(e_1 + e_2)$. The analysis of the employee's incentives alone thus establishes that $ß_1$ must equal $ß_2$ if each tasks is to receive some attention.

APPLICATION: COST CENTERS AND PROFIT CENTERS As the models make clear, an important part of the problem of designing incentives is to determine what the employee will be responsible for, that is, what measures will be used to evaluate performance as a basis for compensation. As an example, consider the problem of providing incentives to the manager of a manufacturing facility. One approach might declare that the manager is responsible only for the costs incurred in the factory, on the theory that the manager has little control over revenues. In that case, we say that the factory is a *cost center*, and the accounting systems should be set up to assess accurately the costs attributable to the factory. Another approach declares that product quality and speed of delivery are important to sales, so that it is a mistake to encourage the manager to focus on cost control at the expense of these factors. Thus, sales performance might be given some weight in determining the manager's compensation.

To represent these issues in terms of our theory, suppose that the two activities to which the manager might contribute are cost reduction and revenue generation. If sales revenues are subject to random variations that are outside the manager's control and statistically independent of the randomness that affects costs, then the cost of providing incentives of strength ß to the manager for revenue generation is the risk premium: $\tfrac{1}{2} rß^2Var(\text{Revenues})$. The equal compensation principle implies that if the factory manager is to be provided with sales-generation incentives at all, then it is futile to do that in a half-hearted way: The incentives need to be of the same strength as those for manufacturing cost control. If the ß associated with cost control is to be large, then the ß associated with revenue generation must be large as well, and

therefore quite costly (in the sense of its leading to a large risk premium). Then, the factory is a *profit center*, to which revenues and costs are both attributed in determining performance.

Cost centers and profit centers are not the only alternatives, however, nor is either likely to be the best alternative in the situation we have described. The firm should actively seek ways to make production managers responsible for what they each control without making them responsible for the performance of the sales force, which they do not control. For example, if quality control and delays in the factory are the chief concerns, then the firm could devise new measures of manufacturing performance, such as the average time from order to delivery and the number of products returned as unsatisfactory. According to the informativeness principle, these measures are superior to measures based on dollar sales because they provide a more informative assessment of the manufacturing manager's actual contribution to the sales effort. As we observed earlier, the firm gains most by improving the measurement of variables that figure most heavily as a basis for compensation.

The equal compensation principle suggests another possibility as well: The manager could be paid a salary with no explicit incentive component. This would be a plausible course of action when manufacturing quality control is important but hard to assess accurately. Of course, the manager will still understand that promotions and pay increases will depend on how superiors assess his or her performance, but at least this solution avoids the distortions in allocation of time and effort and the randomness in compensation brought about by an incentive compensation plan based on arbitrary measures of performance.

This analysis of cost and profit centers focuses only on the issue of compensation. Before leaving this example, however, it is helpful to recall that the actual organization design problem is more involved than that. Managers who are given responsibility for profits, for example, are commonly given broader decision authority than those responsible just for costs or sales. Determining a manager's compensation amounts to deciding what he or she is responsible for, and that decision should be made together with decisions about the scope of the manager's authority.

APPLICATION: INCENTIVES FOR TEACHERS The equal compensation principle can be applied to the recent public policy debate about whether it would be helpful to provide cash incentives for teachers to improve elementary and secondary education. Proponents of cash incentives argue that they would be helpful in focusing teachers on their tasks and motivating them to be innovative in the search for effective ways to train their students.

Opponents of the incentives for teachers, however, have a cogent response. The measures that have been used in the past to evaluate teaching performance for elementary school age children are tests of basic skills, and teaching these is just one part of a teacher's job. Children are also expected to learn social skills, oral expression, and creative thinking, and to build confidence that prepares them for the harder challenges to be faced in later years. Teachers who are compensated based on tests of basic skills alone would be tempted to neglect these other aspects of the job. They might also be led to teaching the most docile students, whose performance scores are easiest to improve, while neglecting students who have more trouble learning. In one instance in South Carolina in 1989, a teacher was caught teaching the answers to the actual test, a copy of which had been illicitly obtained. Compensating teachers based on test scores motivates teachers to help students test well, rather than to help students learn.

According to the equal compensation principle, if it is desirable to have teachers devote some efforts to each of several activities and if it is impossible to distinguish efforts on the various different activities, then all these kinds of efforts must be compensated equally. If social development, oral expression, or creative thinking cannot be accurately measured, then the only realistic options are to remove the responsibility for teaching them from the teacher or to pay the teacher a fixed wage, with no element of incentives pay.

It is a good idea to remember that responsibilities and compensation should really be determined together. In the case of teachers, for example, one proposal is to install a system of specialist teachers who are compensated based on student test scores but who are not responsible for other aspects of student performance.[7] This would not, by itself, solve all the potential problems we have described, but it would allow performance incentives and still ensure that attention is paid to developing the very important "higher thinking skills" in young students. The general point to remember is that by determining the job design and the compensation together, one can sometimes solve problems that cannot be solved by compensation policy alone.

APPLICATION: ASSET OWNERSHIP The equal compensation principle also makes it possible to give a careful treatment of some important issues in the theory of employment and asset ownership. We represent ownership by supposing that at the end of a period of production, the owner of the asset may take it and employ it in other uses. For example, if the employer is the owner of a machine (the asset), he or she can assign the job of production and the use of the asset to another worker, whereas if the worker owns the asset then he or she can employ it on his or her own behalf or on behalf of another employer. What kind of incentives are optimal and who should own the asset?

Assets are notoriously hard to evaluate accurately and objectively. That is why accountants generally report adjusted historical cost figures for asset valuations rather than attempting to account for asset values on the basis of the asset's physical condition (unless deterioration is obvious), its fair market value, or its productivity. The value of a business automobile, for example, is accounted for by its purchase price less an allowance for depreciation, even though its actual value depends on its mileage, physical condition, and so on. Production machines are accounted for in a similar way, even if hard use or changes in production methods has made their actual value lower.

We represent the idea that assets are hard to value accurately in our model by the following assumption: Although the *actual* value of the asset, $A(e_1) + x_1$, is an increasing function of the effort $e_1$ that the worker devotes to maintaining and improving the asset (and of random factors $x_1$), accounting measures of asset values do not reflect those efforts and so cannot be used to provide incentives. Only the direct output of the production process, which is $e_2 + x_2$, is observed by the parties and can serve as a basis for compensation. Therefore, we may write the compensation paid to the worker in the form $\alpha + ß(e_2 + x_2)$.

Suppose that there is some level of total effort $\bar{e}$ that the employee is willing to provide even in the absence of any cash incentives, although this level might be lower than the employer would like to see provided. The efforts $e_1$ and $e_2$ devoted to each of the two activities cannot be observed, however. Should the firm induce greater effort by setting ß positive and thereby inducing more production effort $e_2$?

If the firm owns the asset, then the worker's certain equivalent compensation is $\alpha + ße_2 - \frac{1}{2}ß^2 rVar(x_2) - C(e_1 + e_2)$. If ß is positive, the worker's optimal choice of $e_1$ is always zero. This is just an application of the equal compensation principle:

[7] Jane Hannaway, "Higher Order Skills, Job Design, and Incentives: An Analysis and Proposal," working paper, Stanford University, 1991.

Because the marginal return to the agent from efforts devoted to maintaining or increasing the asset's value is always zero, the worker will devote no efforts to that activity ($e_1 = 0$) unless the returns to other activities are also zero. When the worker is an employee and maintenance of the asset is important, we find (in this model) that it is optimal to pay a fixed wage with no incentives for output performance ($ß = 0$). Then the worker will set $e_1 + e_2 = \bar{e}$ and presumably will be willing to allocate this total amount of effort as the firm directs.

The other possibility is that the worker may own the asset. In that case, the worker's certain equivalent compensation is the sum of the asset's expected value (which depends on $e_1$) and his or her expected compensation, less a risk premium that reflects both the uncertainty in the asset's value and that in the worker's pay, as well as the cost of effort: $A(e_1) + \alpha + ße_2 - \frac{1}{2}rVar(x_1 + ßx_2) - C(e_1 + e_2)$. As the owner, the worker has a built-in incentive to care for the asset; he or she keeps any value that is created when the asset is well cared for. In order to motivate the worker also to pay some attention to production, it is necessary to set $ß > 0$. Then, with positive returns to both types of effort, the worker will choose to provide more total effort than $\bar{e}$—the amount he or she would provide as an employee with no pay incentives for working harder.

To summarize, if it is important that time and effort be devoted to both producing and maintaining the asset, then incentive pay should always be used for workers who bring their own tools ("independent contractors"), but it should never be used for those who use the firm's tools ("employees"). In practice, incentives are used more extensively for independent contractors than for individual employees, as our analysis suggests they should be. The analysis also suggests that independent contractors will work harder than employees, devoting more effort both to caring for the asset and to being directly productive. They will also earn a higher average income to compensate for the extra work they do and the greater risk they bear.

Finally, we come to the question: Who should own the asset? A detailed study of asset ownership is contained in Chapter 9, so we are brief here. In the model just described, if the worker owns the asset, then the worker bears risk both from the randomness of asset returns and from the errors in performance measurement, which add $\frac{1}{2}r(ß^2Var(x_2) + Var(x_1))$ to the total risk premium. Against this must be weighed the fact that the ownership of the asset and increased incentives for the production activity will elicit a higher level of effort. A cost-benefit calculation that balances these considerations must be done to determine which arrangement is likely to be more successful. Certain general principles are evident, however. Increases in the worker's risk aversion or in the variance of asset returns or in the variance of performance estimates in the production task all add to the risk premium that is incurred when the employee owns the asset, making the ownership solution less valuable. If there are many ways to improve performance, then the employee's efforts are especially likely to be responsive to incentives (represented in our model by the assumption that $C''$ is small). Increases in the worker's scope for action tend to favor having the worker own the asset. As we see later, there are a number of other considerations involved in assigning asset ownership efficiently that are not represented in this simple conceptual model.

## Intertemporal Incentives: The Ratchet Effect

An especially thorny problem in real incentive systems is how to set the standard by which performance is to be evaluated. In terms of our model, this means that the mean of $x$ is unknown, so that the mean measured performance corresponding to any fixed level of effort in any performance measurement period is uncertain. If the estimated mean value of $x$ is $\bar{x}$, then the expected level of performance is $e + \bar{x}$ and the corresponding expected pay is $E = \alpha + ß(e + \bar{x})$. If the intended expected compensation is $E$, then the value of $\alpha$ that will be set is determined by rearranging that equation to obtain: $\alpha = E - ß(e + \bar{x})$. Increasing the estimated value of $\bar{x}$ therefore leads to changes in the fixed component of the incentive formula. The magnitude of the effect is proportional to the incentive intensity ß. Setting the standard too high will lead to consistently low levels of pay, perhaps leading to low morale and quits. Setting it too low will lead to happier employees but also to consistently higher levels of pay and lower net profits than would otherwise be possible.

There are just three reasonably objective ways to set a performance standard. The first, which is used frequently only for routine clerical or production tasks, is to have engineers perform *time-and-motion studies* to determine, in detail, how a certain operation is most efficiently done and how long it should take. For example, an engineer might use a stopwatch to determine how long it takes a microfilm operator working at normal speed to load film into the machine, process documents through it, and rewind, catalog, and store the film in the appropriate area. Conducting such studies is costly, and the studies themselves can become obsolete quickly if the job is one where workers can learn and adopt new techniques as they accumulate experience. The second way is to use the performance of other people in similar jobs, that is, to use *comparative performance evaluations*, which we analyzed earlier in the chapter. The third way is to use the *past performance* of the same person in the same job. However, basing standards on past performance penalizes good performance and rewards bad. If workers foresee this possibility, very negative consequences can emerge.

The tendency for performance standards to increase after a period of good performance is called the **ratchet effect**. The term was originally coined by students of the Soviet economic system, who observed that managers of Soviet enterprises were commonly "punished" for good performance by having higher standards set in the next year's plan or, even worse, in the next quarter's plan (see Chapter 1). There are widely known instances of Soviet factory managers who responded to newly installed incentives with massive gains in productivity, only to be denounced on the grounds that their improved performance was proof that they had previously been lazy or corrupt. The ratchet phenomenon is even much older than this example: In a traditional interpretation of the Jewish Passover story, the Egyptian Pharaoh held a brick-producing contest among the Hebrew slaves and used the results as a standard to set a much higher daily production quota.

THE IMPACT OF THE RATCHET EFFECT The "ratcheting up" of standards in response to good performance is not merely unfair, it can be unproductive: If workers foresee the way future standards will depend on current performance, they may refuse to cooperate with efforts to improve productivity. Soviet managers, who are well aware of the ratchet effect, have often been reluctant to institute changes that could radically reduce costs, despite promised incentive payments. Similarly, in the United States, the traditional animosity of labor unions toward piece rates may be explained by the concern that employers, once they have discovered the rates at which highly motivated employees can work, will set a higher standard, leading to lower average pay or to the same average pay for harder work.

There are situations in which there can be efficiency reasons for basing current standards on past performance, but these arise when there are different employees doing the work in different periods. According to the informativeness principle, it is desirable to use all available information that might reduce the variance in the measurement of second-period performance. First-period performance will often give useful information of this sort. The issues here are really identical to those that arise

in the case of comparative performance evaluation, though in this case the comparison is across time periods in the same job rather than across workers at the same time in different jobs.

Using information from past performance lowers the variance with which second-period effort is measured. According to the incentive-intensity principle, the parties will therefore tend to set the incentive term, ß, higher in the second period than they otherwise would, taking advantage of the reduced variance. The theory thus predicts that when the parties write contracts for one period at a time and when past performance embodies useful information for evaluating future performance, incentives will become more intense over time, as the parties utilize past experience to incorporate more accurate performance expectations in their contracts. Because higher levels of ß induce higher levels of effort, the actual effort levels elicited from the worker will also rise over time.

The argument using the informativeness principle is only correct, however, if there is a new occupant in the job in each period. If the same worker is to do the job in each period, then the parties would be better off if they could commit initially to a contract that elicits the same level of effort in each period. This point is subtle, and establishing it will necessitate use of the formal model. The essential source of the inefficiency, however, is that if the worker anticipates that standards in future accounting periods will depend on past performance, then it becomes more difficult or costly to provide him or her with incentives to perform well in the early accounting periods. The example of the Soviet factory managers whose incentives were destroyed by the fear of increased performance standards illustrates the underlying logic of this argument. Because the problem could, in principle, be avoided if the parties could commit themselves in advance not to rely too heavily on past performance for standard setting, we classify it as a problem of *imperfect commitment*.

### *The Mathematics of the Ratchet Effect*

To study the ratchet effect more closely, let us suppose that an employee works for two periods and exerts effort $e_1$ in the first period and effort $e_2$ in the second, but that the contract that is used in the second period is the one that appears optimal at that time, given the available information. Suppose in addition that each of these effort levels is observed only imperfectly: The employer observes only $z_1 = e_1 + x_1$ in the first period and $z_2 = e_2 + x_2$ in the second, where the observation errors in the two periods are assumed to have equal variances and to have means equal to zero. We may write the employee's incentive pay in the first period as $\alpha_1 + ß(e_1 + x_1)$, that is, a constant component $\alpha_1$ plus an incentive component that is proportional to an unbiased estimate of the employee's effort.

It is reasonable to suppose there is a positive correlation between $x_1$ and $x_2$: A high level of $x_1$ means that a high value of $x_2$ is likely. This would occur if some of the same factors that contribute to a high level of measured performance in the first period also contribute to high measured performance in the second. Then the parties can use the observed performance in the first period to get an estimate $\hat{x}_2$ of $x_2$, the part of second-period performance that is beyond the employee's control. This estimate of $x_2$ can then be used to get a better estimate of the employee's actual effort in the second period. That is, the parties may define a standard $\hat{x}_2 = \gamma + \delta(e_1 + x_1)$ and form an adjusted estimate of the employee's second-period performance of the form $\hat{z}_2 = z_2 - \hat{x}_2 = e_2 + x_2 - \hat{x}_2$.

What makes this seem beneficial is that if $\delta$ and $\gamma$ are chosen well, then $Var(x_2 - \hat{x}_2)$ is less than $Var(x_2)$: The adjusted measure eliminates part of the performance variation that is beyond the worker's control and provides a more accurate portrayal of his or her actual performance. Then the informativeness principle indicates that this estimate should be used in the contract. For example, if a retail chain bases each store manager's pay partly on the level of sales at the store, then it seems only fair (and efficient) that those stores in good locations should have to meet a higher sales target. Surely, there is no more generally accurate way to determine the quality of a location than by looking at the past record of the store's sales, and that is just what the proposed standard achieves.

If the employee's second-period pay is given by the same sort of function as in the first period—a constant term plus an incentive term that rewards higher estimated effort—then the employee's total compensation over the two periods is $\{\alpha_1 + ß_1(e_1 + x_1)\} + \{\alpha_2 + ß_2[e_2 + x_2 - \gamma - \delta(e_1 + x_1)]\}$. Collecting terms allows us to rewrite this as

$$\text{Total Compensation} = \alpha_1 + \alpha_2 + (ß_1 - \delta ß_2)(e_1 + x_1) + ß_2(e_2 + x_2 - \gamma) \tag{7.11}$$

Note, however, that the coefficient of $e_1$ in Equation 7.11 is not the nominal contractual amount, $ß_1$, but rather the smaller amount $ß_1 - \delta ß_2$. This is the ratchet effect at work. The direct return—in terms of first-period pay—to extra effort in the first period is $ß_1$, but higher first-period effort increases the standard in the second period by $\delta$. It thus reduces the pay accruing in the second period for any choice of second-period effort by $\delta ß_2$. If the same employee occupies the job in both periods, then the anticipation of first-period performance being used in the second period reduces effective first-period incentives. Thus, it makes sense to define the "effective incentives" $ß_1^E$ and $ß_2^E$ by $ß_1^E = ß_1 - \delta ß_2$ and $ß_2^E = ß_2$.

In terms of the effective incentives, the total certain equivalent wealth can be written as

$$\text{Total Certain Equivalent} = P(e_1) + P(e_2) - C(e_1) - C(e_2) - \tfrac{1}{2}rVar(ß_1^E x_1 + ß_2^E x_2) \tag{7.12}$$

An efficient contract maximizes this expression subject to the incentive constraints, $ß_1^E = C'(e_1)$ and $ß_2^E = C'(e_2)$, that determine the worker's effort choices. Using properties of the variance (see the formulas in the appendix), Equation 7.12 can be rewritten as

$$\text{Total Certain Equivalent} = P(e_1) + P(e_2) - C(e_1) - C(e_2) - \tfrac{1}{2}r[(ß_1^E)^2 Var(x_1) + (ß_2^E)^2 Var(x_2) + 2ß_1^E ß_2^E Cov(x_1, x_2)] \tag{7.13}$$

where $Cov(x_1, x_2)$ is a measure of the way the two measurement errors tend to move together. Recall that we have assumed that $Var(x_1)$ and $Var(x_2)$ are equal. Both Equation 7.13 and the incentive constraints are symmetrical with respect to time: They would be essentially unchanged if we were to change each $e_1$ into an $e_2$, each $e_2$ into an $e_1$, each $ß_1^E$ into $ß_2^E$, and each $ß_2^E$ into $ß_1^E$. Thus, if there is a unique total wealth maximizing contract, it must treat the two time periods symmetrically. That is, it must have $e_1 = e_2$ and $ß_1^E = ß_2^E$.

In contrast, we saw that when the parties cannot commit in advance to the second-period contract terms and instead act optimally in the second period given what has already transpired, they will set $e_2 > e_1$: Incentives are made more intense over time.

OVERCOMING THE RATCHET EFFECT The parties would be better off if they could commit to hold the line on incentives, not using the first-period performance to adjust second-period performance standards. In that case, the contractual and effective incentives would be the same. This policy is in fact in place at some companies. The Lincoln Electric Company is famous for its extensive use of incentive contracts and, in particular, piece rates. Lincoln has for decades maintained a policy that once a piece rate has been set, it will not be changed unless the equipment is changed or new work methods are introduced. In this case, a new time-and-motion study will be done, and the resulting standard will remain in effect even if realized performance later suggests that it is too low. If the standards are set too low, then the Lincoln workers may earn a lot of extra money, but their incentives to work hard are never threatened.

Why don't more companies adopt such a system? There are many reasons, not least of which is that it is so difficult for a firm to commit itself not to use available information.[8] Even if the parties agree in advance not to use the information embodied in the first-period performance and set contract terms accordingly, there will still be efficiency gains after the fact to renegotiating the contract and using the information. Lincoln Electric has managed to commit to its policy of maintaining standards by applying piece rates widely throughout its organization, developing expertise at doing time-and-motion studies, building a reputation for applying its "no revisions" policy consistently, and tuning the rest of its policies so that they are consistent with a piece-rate system.

Our characterization of the ratchet effect as a problem of commitment also helps us to understand how *self-employment* arrangements and *ownership* can sometimes alleviate the problem. A self-employed person sells goods or services directly to customers. If the industry is competitive, then the performance standards are the comparative ones set by the marketplace, and a person's good performance does not lead directly to higher future standards. Similarly, someone who owns an asset can be assured of keeping whatever gains accrue from showing how to use it effectively. Of course, the problems of risk sharing often make the ownership solution impractical, and self-employment is infeasible in many situations.

Within companies, job rotation is another device that can be used to alleviate the ratchet effect. By assigning people to various jobs over time and using previous jobholders' performance to set the standard, the current jobholder is not penalized for a job well done. Job rotation also may bring benefits in improved morale and greater flexibility in production. Its costs, however, are in potentially reduced efficiency, as workers have less opportunity to gain experience in any task.

## MORAL HAZARD WITH RISK-NEUTRAL AGENTS

The main thrust of this chapter has been to study principal-agent problems where motivating agents by making them bear part of the risk is costly because the agents are risk averse. From an analytical perspective, we can organize our study of other aspects of the theory by assuming away the risk-sharing problem and studying principal-agent theory with a risk-neutral agent, that is, one whose coefficient of absolute risk aversion is equal to zero. In that case, no risk premium is ever incurred, regardless of how the risks are shared. So, the agent can be perfectly motivated at zero cost by setting $\beta = 1$, that is, by making him or her bear the entire risk. For a manager-agent running a firm, this is very much like having the manager buy the firm and enjoy all the profits and suffer all the losses. In the case of automobile insurance, it amounts to having drivers paying full cash compensation to those who they have damaged. There are several reasons, however, why a solution of this kind would often be unworkable, and each of these points to a factor that makes the moral hazard problem more difficult to resolve.

[8] There may be other reasons that have little or nothing to do with incentive issues directly. For example, in a multistep production process, there may be little value to having one worker proceeding faster than the others, so the firm may not want to encourage every individual worker to work as fast as he or she individually can.

### Problems with the Risk-Neutral Agent Scenario

When is making the risk-neutral agent responsible for all financial losses not a workable solution? First, the solution will fail whenever the agent lacks sufficient funds. A manager may simply be unable to guarantee payments for the business's expenses with personal funds, and a driver may be unable to pay for damages from a serious accident. Arrangements in the economic world are often made with these limits in mind. Public policy toward drivers generally requires that automobile owners have insurance or show some other evidence of financial responsibility before their cars can be licensed. In the private sector, vendors are frequently unwilling to extend trade credit to a company with little working capital, for fear that their bills will never be paid.

A second case where making the agent bear the risk is not workable is when the risk is a nonfinancial one and is therefore difficult or impossible to transfer. There is no way for a careless or drunken driver to undo the injuries or death that may result from an automobile accident simply by paying damages, or for a negligent blood bank to bear the suffering of a recipient of AIDS-tainted blood, or for a company that dumps toxic or radioactive waste to eliminate the genetic damage done to victims simply by paying a cash penalty. All of these examples pose important problems for public policy, but the principles that arise are not the kind that are most helpful for understanding the institutions and practices of the business world. So we merely note that these problems do limit the theory and pass on.

ADVERSE SELECTION IN THE PRINCIPAL-AGENT PROBLEM A third case in which "selling the firm to the manager" is not a workable solution is when the principal and the agent cannot agree on a price, for example, because the market is disrupted by adverse selection. This variation is well typified by the case of an employee of a department store chain who invents a new consumer product. Being no expert at marketing, the employee negotiates with the chain to market the product. What kind of arrangements should the employee make for marketing the product?

The employee in this case is the principal who is trying to negotiate a contract to motivate the agent (the department store chain) to market the product. Unlike our earlier examples, this is a case in which the agent is much more tolerant of risk than is the principal, so efficient risk sharing would dictate that the agent bear (almost) all the risk, and efficient incentives for marketing effort by the agent would seem to lead to the same conclusion. If these were the only factors involved, the efficient solution would be to sell the rights for the product to the department store chain, which would then bear all the market risk. Furthermore, as the owner the chain would be motivated to work as hard as necessary to extract all the potential value from the product.

However, there is another element that may block this easy solution—the element of adverse selection. In this case, the chain is an expert marketer who may be much better informed than is the employee/inventor about the market potential of the product. If the rights are to be sold to the chain, how will the price be determined? The chain will refuse the principal's offer whenever its estimate of sales is low and will accept the offer when its estimate is high. Therefore, the inventor can only successfully sell the rights for a price that is lower than the expected profitability of the product.

An alternative procedure would be for the inventor to keep the right to the

product and demand royalties from the chain, that is, a payment proportional to the number or value of the units sold. In that way, the inventor can mitigate the adverse selection problem because the employer could be expected to accept a royalty contract regardless of its private information about the sales potential of the invention. It only has to pay an amount proportional to its actual sales. There are two problems with this option too, however. First, the inventor is made to bear too much risk, and, second, the department store chain, which no longer receives all the profits from sales, will be inclined to expend too little effort promoting sales of the product.

Another way that the employee/inventor might try to avoid the problem of distorting the chain's incentives while receiving some royalties is to base the royalties on profits rather than on sales. In that way, the chain would be motivated to incur the right amount of costs to maximize profits because its share of cost is the same as its share of the revenues. The drawback of this scheme is that the accounting for expenses is in the hands of the chain, and the employee/inventor ought to expect that the accounts will be manipulated to reduce royalty payments. Indeed, in the 1980s and in 1990, there were well-publicized lawsuits by film makers and actors against Hollywood studios that had agreed to pay a percentage of profits on movies or television shows, only to claim that there was little to share. In a celebrated case, actor James Garner sued over his rights to royalties from *The Rockford Files*, one of the most successful television shows in history. According to the accounting procedures used by the studio, however, the show never earned a profit.

The formal analysis of efficient contracting when there is both moral hazard and adverse selection is quite complex. The inventor's best policy in the situation we describe depends on his or her risk aversion, on the importance of motivating the chain to promote the product aggressively and how that importance depends on circumstances, and on the quality of the chain's sales forecasts, among other variables. Although the theory has little to say about the *details* of the solution, it has much to say about the form. In a broad range of cases, the best the inventor can do is to offer the chain a choice between purchasing the full rights to the invention at a relatively high price or paying a lower price plus royalties proportional to sales. The actual prices and royalties will depend on the parties' relative bargaining power, but the inventor should anticipate during the negotiations that the chain will want to own the rights when it forecasts high sales and to pay a royalty when its forecasts are less optimistic. If the chain insists that the rights are not worth much, then the inventor should insist on receiving royalties instead of a fixed payment, even though this may cause the chain to promote the product less effectively. By selling the invention outright when its value is high, the inventor motivates the chain to promote sales vigorously in that case and so increases the value of the invention.

The case of the inventor and the chain store is important because the problems that arise and the pattern of analysis are similar to those in many other business settings. The key characteristic of these settings is that there is one party that has superior information about costs and that needs to be motivated to work hard. For example, in setting procurement policy, governments (and firms) depend on suppliers to supply appropriate information about costs and recommendations about product design and also to work hard to build quality into the products supplied. In the regulation of utilities, the public utility commissions rely on the regulated firms both to supply information that will be the basis for price decisions and to work hard to keep costs as low as possible.

Do all these variations invalidate the general principles presented earlier in the chapter? They do not. Although such principles as the informativeness principle and the equal compensation principle are derived from and phrased in terms of a particular conceptual model (in which agents are risk averse and must be motivated to undertake personally costly effort), both can be rephrased to hold over the whole range of variations described here. Of course, the general principles cannot, by themselves, substitute for analysis of particular cases, but they do provide a useful guide across a wide range of applications.

## SUMMARY

Most people in the economy dislike bearing risk. The cost of risk bearing can often be reduced by sharing the risk among a group of people. If the group is large and the risks that different people contribute are statistically independent, this procedure can virtually eliminate the cost of bearing these risks. Insurance companies exist primarily to perform this economic function. Some kinds of risks, however, are not easily insured, principally because they are risks that affect many people simultaneously (such as environmental risks and energy shortages) and so would threaten the capital of any insurance company. These kinds of risks are managed and shared through other institutions, with the securities markets playing a prominent role when the risks are expressible in money terms.

Principal-agent problems are situations in which one party (the principal) relies on another (the agent) to do work or provide services on his or her behalf. When agents' actions cannot be easily monitored and their reports easily verified, the agents have greater scope to pursue their own interests rather than the principal's. Then, to provide incentives for the agents to behave in the principal's interests, it is necessary to arrange for them to bear some responsibility for the outcomes of their actions and therefore to bear more risk than would otherwise be desirable.

Several principles govern the design of optimal incentive contracts. The *informativeness principle* says that the cost of providing incentives increases with the variance of the estimator of the employee's effort. An optimal incentive contract should base the employee's compensation (or the insured's contribution to cover a loss) *only* on the minimum variance indicator of the employee's behavior. This principle is applied to illuminate issues of comparative performance evaluation and the use of deductibles in automobile insurance and copayments in health insurance.

The *incentive-intensity principle* says that the strength of incentives should be an increasing function of the marginal returns to the task, the accuracy with which performance is measured, the responsiveness of the agent's efforts to incentives, and the agent's risk tolerance. This principle is useful for explaining variations in the strength of incentives among Japanese subcontractors and among general partners in certain oil and gas drilling programs.

The *monitoring intensity principle* states that more resources should be spent monitoring when it is desirable to give strong incentives. This principle is the mirror image of the observation that more accurate performance information leads to higher optimal incentives, which is part of the incentive intensity principle. Measuring performance carefully and providing intense incentives are complementary activities, which should be found together.

The *equal compensation principle* holds that if an employee's allocation of time and effort between alternative tasks cannot be monitored by the employer, then the marginal returns earned by the employee in any tasks to which he or she actually devotes effort must be equal. Providing strong incentives for a portion of an employee's activities can cause the employee to cut back his or her efforts in other activities. This principle informs comparisons of cost centers and profit centers, policy debates about the provision of incentives for elementary and secondary school teachers, and analyses of asset ownership patterns.

The *ratchet effect* refers to the practice of basing performance targets on past performance in the same activity. Although such a practice would seem to be consonant with the informativeness principle because it bases performance goals on one of the best available indicators of possible performance, it also imposes certain costs. If the occupant of the job does not change over time, then the effect of the ratchet is to punish yesterday's good performance by setting higher standards today. In modern capitalist economies, ownership and self-employment are two principal means by which a person who performs well can guard against being subjected to the ratchet effect.

There are various factors not included in our models that would make the principal-agent problem more difficult to resolve. One is that the agent may lack sufficient capital to pay penalties for losses that he or she causes. A second is that some losses are essentially nonfinancial losses that cannot be easily compensated using cash. A third is that the agent may have private information that makes it more difficult for the principal and the agent to agree on contract terms.

## BIBLIOGRAPHIC NOTES

The economic theory of decisions under uncertainty has its origins in Daniel Bernoulli's eighteenth-century writings. This theory was put on a sound logical foundation in the 1940s through the collaborative efforts of the mathematician John von Neumann and the economist Oskar Morgenstern. A modern treatment of this theory has been presented by David Kreps. The refinements contributed by Kenneth Arrow and John Pratt (and reviewed in the appendix) made it possible to apply the von Neumann and Morgenstern theory to analyze risks that are specifically financial in nature. Their work led to the formulas for risk premia and certain equivalents that we have used. The theory of risk sharing and insurance was built on these foundations in the 1960s by Karl Borch and Robert Wilson.

The modern theory of incentives was begun in the 1970s by various authors who explored what optimal incentive compensation contracts might be like in a variety of different applications including: insurance (Michael Spence and Richard Zeckhauser), sharecropping (Joseph Stiglitz), tax policies (James Mirrlees), and managerial compensation (Robert Wilson and Stephen Ross). Stephen Shavell and Bengt Holmstrom originated the informativeness principle, which was later generalized in the work of Sanford Grossman and Oliver Hart. Milton Harris and Artur Raviv also made early important contributions to understanding the nature of efficient contracts. The ratchet effect has been analyzed by Martin Weitzman, David Baron and David Besanko, and Xavier Freixas, Roger Guesnerie and Jean Tirole. The problem of renegotiation of principal-agent contracts was first studied by Mathias Dewatripont; see also the contributions by Philippe Aghion, Dewatripont, and Patrick Rey and by Drew Fudenberg and Jean Tirole. Our discussions of optimal linear incentive contracts and the other principles of incentive pay borrow heavily from the work of Bengt Holmstrom and Paul Milgrom.

There were many contributions to incentive theory in the 1980s focusing on the case where there is a tension between the needs to alleviate adverse selection and moral hazard. In addition to a number of those listed above, leading contributors to that theory were Joel Demski, Jean-Jacques Laffont, Preston McAfee, John McMillan, Roger Myerson, Michael Riordan, and David Sappington. Their theories were often set in the particular situation of a government regulator (principal) trying to regulate a utility (agent), or a procurement officer (principal) trying to negotiate a complex contract with a supplier (agent). The principles that have emerged from these analyses, however, have wide application.

## References


Aghion, P., M. Dewatripont, and P. Rey. "Renegotiation Design under Symmetric Information," mimeo, 1989.

Arrow, K. J. *Essays in the Theory of Risk Bearing* (Chicago: Markham, 1970).

Baron, D., and D. Besanko. "Regulation and Information in a Continuing Relationship," *Information, Economics and Policy*, 1 (1984), 267–330.

Baron, D., and R. Myerson. "Regulating a Monopolist with Unknown Costs," *Econometrica*, 50 (July 1982), 911–30.

Bernoulli, E. "Specimen theoriae novae de mensura sortis," *Commentarii Academiae Scientiarum Imperialis Petrepolitanae*," (trans) "Exposition of a New Theory on the Measurement of Risk," *Econometrica*, 22 (January 1954), 23–36.

Borch, K. "Equilibrium in a Reinsurance Market," *Econometrica*, 30 (July 1962), 424–44.

Demski, J., and D. Sappington. "Optimal Incentive Contracts with Multiple Agents," *Journal of Economic Theory*, 33 (1984) 152–71.

Dewatripont, M. "Renegotiation and Information Revelation over Time in Optimal Labor Contracts," *Quarterly Journal of Economics*, 104 (1989), 589–620.

Freixas, X., R. Guesnerie, and J. Tirole. "Planning Under Incomplete Information and the Ratchet Effect," *Review of Economic Studies*, 52 (1985), 173–92.

Fudenberg, D., and J. Tirole. "Moral Hazard and Renegotiation in Agency Contracts," *Econometrica*, 58 (November 1990), 1279–1320.

Grossman, S., and O. Hart. "An Analysis of the Principal-Agent Problem," *Econometrica*, 51 (1983), 7–45.

Harris, M., and A. Raviv. "Optimal Incentive Contracts with Imperfect Information," *Journal of Economic Theory*, 20 (1979), 231–59.

Holmstrom, B. "Moral Hazard and Observability," *Bell Journal of Economics*, 10 (1979), 74–91.

Holmstrom, B. "Moral Hazard in Teams," *Bell Journal of Economics*, 13 (1982), 324–40.

Holmstrom, B., and P. Milgrom. "Aggregation and Linearity in the Provision of Intertemporal Incentives, *Econometrica*, 55 (March 1987), 303–28.

Holmstrom, B., and P. Milgrom. "Multi-task Principal-Agent Analysis: Incentive Contracts, Asset Ownership and Job Design," SITE Working Paper #6, Stanford University, 1990.

Kreps, D. *Notes on the Theory of Choice* (Boulder, CO: Westview Press, 1988).

Laffont, J. J., and J. Tirole. "Using Cost Observations to Regulate Firms," *Journal of Political Economy*, 94 (June 1986), 614–41.

Laffont, J. J., and J. Tirole. "The Dynamics of Incentive Contracts," *Econometrica*, 56 (1986), 1153–75.

McAfee, R. P., and J. McMillan. "Competition for Agency Contracts," *Rand Journal of Economics*, 18 (1987), 396–7.

Mirrlees, J. "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 38 (1971), 175–208.

Mirrlees, J. "Notes on Welfare Economics, Information, and Uncertainty," in *Essays on Economic Behavior Under Uncertainty*, M. Balch, D. McFadden, S. Wu, eds. (Amsterdam: North-Holland Publishing Co., 1974).

Mirrlees, J. "The Optimal Structure of Incentives and Authority within an Organization," *Bell Journal of Economics*, 7 (1976), 105–31.

Pratt, J., "Risk Aversion in the Small and in the Large," *Econometrica*, 32 (1964), 122–36.

Riordan, M., and D. Sappington. "Information, Incentives and Organizational Mode," *Quarterly Journal of Economics*, 102 (1987), 243–64.

Ross, S. "The Economic Theory of Agency: The Principal's Problem," *American Economic Review*, 63 (1973), 134–39.

Shavell, S. "Risk Sharing and Incentives in the Principal and Agent Relationship," *Bell Journal of Economics*, 10 (1979), 55–73.

Spence, A. M., and R. Zeckhauser. "Insurance, Information and Individual Action," *American Economic Review*, 61 (1971), 380–87.

Stiglitz, J. "Incentives and Risk Sharing in Sharecropping," *Review of Economic Studies*, 64 (1974), 219–56.

Stiglitz, J. Incentives, Risk and Information: Notes Towards a Theory of Hierarchy," *Bell Journal of Law, Economics and Organization*, 6 (1975), 552–79.

von Neumann, J., and O. Morgenstern, *The Theory of Games and Economic Behavior*, (Princeton: Princeton University Press, 1944).

Weitzman, M. "The Ratchet Principle and Performance Incentives," *Bell Journal of Economics*, 11 (1980), 302–8.

Wilson, R. "The Theory of Syndicates," *Econometrica*, 36 (January 1968), 119–32.

Wilson, R. "The Structure of Incentives for Decentralization," in *La Decision* (Paris: Centre Nationale de la Recherche Scientifique, 1969).


## Exercises

### Food for Thought

1. In the late 1960s and early 1970s, when McDonalds (the fast-food chain) was undergoing a period of very rapid expansion in sales, it considered a variety of different compensation systems for its managers. The company wanted to encourage its managers to increase sales, control costs, and maintain the company's standards of quality, service, and cleanliness. It also wanted local store managers to hire and train people who could become managers of new outlets, which were being added to the chain at a rapid pace. What difficulties would you expect this situation to pose for McDonald's management? What would you expect to occur if a local outlet manager's compensation were based primarily on sales growth? On outlet profits? What kind of compensation plan should McDonald's adopt? How would you expect the compensation formula to change as McDonald's moved into its next phase, with fewer new outlets being opened in North America?

2. A common complaint of university students is that professors seem too remote and uninterested in teaching them. How do university systems of compensation, promotion, and tenure contribute to the problem? Is the problem likely to be more severe for tenured or untenured faculty? Why do universities often have rules restricting outside consulting activities?

3. Use Figures 7.3 and 7.4 to determine both how the level of monitoring

and the intensity of incentives would change (1) if the total cost of monitoring were to fall by a fixed amount and (2) if the marginal cost of monitoring were to fall.

4. Unlike specialty stores, department stores sell a wide array of products to a single group of customers, and are often especially interested in maintaining their reputations for servicing their customers well. How might this consideration affect the compensation of department store sales personnel compared to the salespeople at specialty outlets?

5. Suppose a Canadian subsidiary of a British company assembles a product using inputs manufactured in Canada. Under what conditions should the manager of the British firm be responsible if a change in the foreign exchange rate raises the cost in pounds sterling of purchasing the inputs, causing losses to result? Under what conditions should the manager not be held responsible?

6. In 1989 the NBC television network in the United States announced a new way of compensating the local television stations affiliated with the network. These independently-owned stations carry NBC programs during "prime-time" evening hours from 8:00 to 11:00 p.m. NBC earns its revenues by selling advertising time to national advertisers whose ads appear during the network's shows. The amount NBC gets for its advertising time depends on the number of viewers its programs attract. The number of viewers watching a particular station (and thus the network's programs) during prime-time depends in part on the number of people who are attracted to watch the non-network programs that the station shows before prime time. These viewers tend to stay with the channel that they started watching on a given evening. NBC's innovation was to begin paying its affiliates in part on the basis of the number of viewers that they attracted for their early-evening programming, such as local news, rather than just on the size of the audiences that watched the prime-time network programs on the station. Analyze this plan in terms of the principles developed in this chapter.

## Mathematical Exercises

1. Suppose that a group of people A, . . ., Z share an income risk $I$, in proportion to their risk tolerances. For example, individual A bears a share $\rho_A/(\rho_A + \ldots \rho_Z)$ where $\rho_A = 1/r_A$, and so on. Show that A's risk premium in this case is $\frac{1}{2}\rho_A Var(I)/(\rho_A + \ldots \rho_Z)^2$ and hence that the total risk premium borne by all the members is $\frac{1}{2}Var(I)/(\rho_A + \ldots + \rho_Z)$—the same as the risk premium that would be required by a single person whose risk tolerance is $\rho_A + \ldots \rho_Z$.

2. Consider the case of two people, A and B, with incomes $I_A$ and $I_B$, and suppose that they enter a risk sharing contract so that A's income is $\alpha I_A + ßI_B + \gamma$, with B receiving the balance. The total risk premium for this arrangement is given by Equation 7.1 in the chapter. (a) Expand this equation into one expressed in terms of $Var(I_A)$, $Var(I_B)$, and $Cov(I_A, I_B)$. [Hint: Your answer should be a quadratic function of $\alpha$ and ß.] (b) To find the values of $\alpha$, ß, and $\gamma$ that minimize this expression, take the derivatives of the expression with respect to $\alpha$ and ß and set the derivatives equal to zero. Show that the solution of these three equations has $\alpha = ß = \rho_A/(\rho_A + \rho_B)$, where $\rho_A = 1/r_A$ and $\rho_B = 1/r_B$ are the two risk tolerances. (c) Use mathematical induction and the results of this and the preceding problem to show that regardless of the number of people, person 1's share of each risk should be the same as his or her share of the total risk tolerance of the group.

3. In the text we compared the advantages of relative performance evaluation against an evaluation based solely on the employee's own performance. Here we consider all combinations of the two as well. Thus, suppose manager A's measured performance is $e_A + x_A + x_C$ and B's measured performance is $e_B + x_B + x_C$, where $x_A$, $x_B$, and $x_C$ are independent sources of randomness. Suppose it is proposed to base manager A's compensation on his or her own performance minus $\delta$ times some measure of B's performance. Find the value of $\delta$ that minimizes the variance of the performance measure. How does this value change with changes in $Var(x_A)$? Changes in $Var(x_B)$? Changes in $Var(x_C)$?

4. Suppose an entrepreneur can select among investment projects that all cost the same amount but differ in their risk-return characteristics. The set of available projects is described by a curve giving the highest available expected net return (after subtracting the initial cost of the investment) corresponding to any given variance in the returns. Let this curve be $m = 2v - (\frac{1}{2})v^2$, $0 \leq v \leq 2$, where $m$ is the mean return and $v$ is the variance of returns. Thus, the entrepreneur can achieve a riskless return of $m = \$0$, essentially by not investing, while the maximum expected return is attained by selecting a project with $v = 2$, which yields an expected return of 2. What project will the entrepreneur choose if he or she must bear all the returns (positive or negative) alone and he or she has a coefficient of risk aversion of $r > 0$, so that his or her preferences are given in certain-equivalent form by $m - (\frac{1}{2})rv$?

Now suppose that it is possible to share the risk of the investment with an outside investor who has a coefficient of risk aversion of $s$. What is the investment choice that maximizes the total certain equivalent? How should the risk be shared? Could this be achieved by selling an ownership claim in the entrepreneur's firm to the investor in such a way that the investor will be willing to pay enough that the entrepreneur is better off selling this share?

5. A risk-averse entrepreneur is considering selling stock in his or her company to the public. He or she will continue to manage the firm after it "goes public." The entrepreneur gets utility from income, $x$, and from the consumption of on-the-job perquisites, $c$, according to the utility function $u(x, c) = \bar{x} - \frac{1}{2}\text{var}(x) + 100c^{\frac{1}{2}}$, where $\bar{x}$ is the mean of the income $x$ and var($x$) is its variance. The uncertain profits of the firm are $Y - c$: each dollar spent on perquisites reduces profit by a dollar. The variance of Y, (and thus of $(Y - c)$), is $\sigma^2$, which we assume to be greater than 2,500. The entrepreneur is currently the sole owner of the firm, receiving as income the firm's profit. What level of perquisites will he or she choose?

Now, suppose the entrepreneur sells a fraction $\alpha$ of the firm to risk-neutral investors, retaining $(1 - \alpha)$ for him- or herself. Thus, the entrepreneur receives as income whatever amount the investors pay for this fraction of the firm, say M($\alpha$), and then gets his or her share, $(1 - \alpha)(Y - c)$, of the random profit. The variance of his or her income is thus $(1 - \alpha)^2\sigma^2$. What is the relationship between $\alpha$, the fraction of ownership the entrepreneur sells, and the level of $c$ he or she will subsequently choose? Does this choice maximize total value (the expected utility of the entrepreneur plus that of the investors)? What will be the expected profit as a function of $\alpha$?

Assume that competition among investors leads them to pay an amount for any given ownership share equal to the profits they expect to receive. How much will the entrepreneur receive from selling a fraction $\alpha$ of the firm if investors correctly anticipate the level of $c$ that the entrepreneur will choose after selling that fraction of the firm? What is the realized level of expected utility for the entrepreneur from selling a fraction $\alpha$ of the firm if the investors have correct expectations? What is the best level of $\alpha$ for him or her to pick? Could the entrepreneur gain by binding him- or herself not to increase c as $\alpha$ changes?

# MATHEMATICAL APPENDIX

This appendix consists of two parts. The first is a review of statistical concepts. The second derives the approximation reported in the main text regarding the certain income that is equivalent, from the decision maker's point of view, to a given random (uncertain) income.

## REVIEW OF STATISTICAL CONCEPTS

The set of possible outcomes in a statistical problem is represented by a **sample space** S. A **random variable** $x$ is a function that associates with each element $s \in S$ a real number $x(s)$. For example, the elements of S may be the books on a bookshelf and $x(s)$ may specify the number of pages in the book $s$. In a coin-tossing problem, the elements of S may be sequences of heads and tails and $x$ may be some statistic, such as the one that assigns to each sequence $s \in S$ the number of heads that occur in the first ten coin tosses. With the sample space S comes a **probability mass function** $p$ that assigns a probability $p(s)$ to each element of $s \in S$. The probability mass function is used to compute such probabilities as $Prob(x = 11)$—the probability that the random variable $x$ takes the value 11.

Each random variable $x$ has a *mean* denoted by $\bar{x}$, also called its **expectation** and denoted $E[x]$. The formula for calculating expectations is:

$$E[x] = \Sigma_{s \in S} p(s)x(s) = \bar{x} \tag{7.14}$$

Each random variable also has a *variance* given by:

$$Var(x) = E[(x - \bar{x})^2] = \Sigma_{s \in S} p(s)(x - \bar{x})^2 \tag{7.15}$$

Variance is one measure (among many possible measures) of the degree of randomness of $x$.

Given two random variables $x$ and $y$, the **covariance** of $x$ and $y$ is:

$$Cov(x, y) = E[(x - \bar{x})(y - \bar{y})] \tag{7.16}$$

Notice that $Cov(x, x) = Var(x)$.

Given two random variables $x$ and $y$ and two real numbers $\alpha$ and $\beta$, we can form a new random variable $\alpha x + \beta y$, which for each possible outcome $s$ takes the value $\alpha x(s) + \beta y(s)$. Its expectation can be computed from those of $x$ and $y$ by the following formula, which can be derived from Equation 7.14:

$$E[\alpha x + \beta y] = \alpha E[x] + \beta E[y] \tag{7.17}$$

Using Equations 7.14–7.17, we can derive the formula:

$$Var(\alpha x + \beta y) = \alpha^2 Var(x) + \beta^2 Var(y) + 2\alpha\beta Cov(x, y) \tag{7.18}$$

The two random variables $x$ and $y$ are *(statistically) independent* if for all numbers $\alpha$ and $\beta$, $Prob[x = \alpha \text{ and } y = \beta] = Prob[x = \alpha] \cdot Prob[y = \beta]$. Statistical independence represents the idea that knowing the value of one of the variables provides no information about the value of the other. If $x$ and $y$ are independent, then $Cov(x,y) = 0$, so by Equation 7.18, $Var(x+y) = Var(x) + Var(y)$.

## EVALUATING FINANCIAL RISKS: CERTAIN EQUIVALENTS AND RISK PREMIA

Expected utility theory establishes conditions under which a decision maker will rank risky prospects according to their associated expected utilities. Let $u$ be a function that assigns to each monetary outcome $x$ a utility $u(x)$. Then, representing prospects by random variables, the expected utility of prospect $x$ is $E[u(x)]$. Let us compare this prospect with a certain prospect, that is, one that yields the payment $\hat{x}$ with probability 1. The certain prospect will be preferred if $u(\hat{x}) > E[u(x)]$; the risky prospect will be preferred if the reverse inequality holds. When the decision maker is just indifferent between the two prospects, then $\hat{x}$ is called the certain equivalent of the prospect $x$.

The crucial formula for our applications is the following approximation:

*Approximation.* Suppose that $u$ is three times continuously differentiable, that $\bar{x} = E[x]$, and that $u'(\cdot) > 0$. Then, approximately, the certain equivalent is:

$$\hat{x} \approx \bar{x} - \tfrac{1}{2} r(\bar{x}) Var(x) \tag{7.19}$$

where $r(\bar{x}) = -u''(\bar{x})/u'(\bar{x})$.

*Derivation.* According to Taylor's theorem, for any $z$,

$$u(z) = u(\bar{x}) + (z - \bar{x})u'(\bar{x}) + \tfrac{1}{2}(z - \bar{x})^2 u''(\bar{x}) + R(z)$$

where $R(z) = u'''(\hat{z})(z - \bar{x})^3/6$ for some $\hat{z} \in [\bar{x}, z]$. We assume that this remainder term is negligible. Hence we write, approximately,

$$u(z) \approx u(\bar{x}) + (z - \bar{x})u'(\bar{x}) + \tfrac{1}{2}(z - \bar{x})^2 u''(\bar{x}) \tag{7.20}$$

Substituting $x$ for $z$ in Equation 7.20 and computing the expectation, we find, approximately,

$$E[u(x)] \approx u(\bar{x}) + E[x - \bar{x}]u'(\bar{x}) + \tfrac{1}{2} E[(x - \bar{x})^2]u''(\bar{x})$$

But, $E[x - \bar{x}] = E[x] - \bar{x} = \bar{x} - \bar{x} = 0$, so approximately,

$$E[u(x)] \approx u(\bar{x}) + \tfrac{1}{2} E[(x - \bar{x})^2]u''(\bar{x}) \tag{7.21}$$

We expect that the certain equivalent $\hat{x}$ will be close to $\bar{x}$, so we approximate its utility differently, also using Taylor's theorem,

$$u(\hat{x}) = u(\bar{x}) + (\hat{x} - \bar{x})u'(\bar{x}) + \dot{R}(\hat{x}) \tag{7.22}$$

where, $\dot{R}(\hat{x}) = \tfrac{1}{2}u''(\hat{z})(\hat{x} - \bar{x})^2$ for some $\hat{z} \in [\bar{x}, \hat{x}]$. Because we apply the approximation only when $\hat{x} - \bar{x}$ is small, we again treat the remainder term as negligible. For a certain equivalent, we have $u(\hat{x}) = E[u(x)]$. So, combining Equations 7.21 and 7.22, we have, approximately,

$$(\hat{x} - \bar{x})u'(\bar{x}) \approx \tfrac{1}{2} E[(x - \bar{x})^2]u''(\bar{x}) \tag{7.23}$$

This may be expressed in the form

$$\hat{x} - \bar{x} \approx \tfrac{1}{2} \cdot [u''(\bar{x})/u'(\bar{x})] \cdot E[(x - \bar{x})^2] = -\tfrac{1}{2} \cdot r(\bar{x}) \cdot Var(x) \tag{7.24}$$

which establishes the Approximation (7.19).