have been Pareto inefficient. More generally, whenever the MRT is not equal to the MRS, such a rearrangement of resources is feasible and hence the optimal product mix condition is

$$MRT = MRS \tag{2.17}$$

Note that it does not matter whose MRS is taken, because all individuals have the same MRS in a market equilibrium.

We now want to show that a competitive market economy achieves an optimal product mix: From the pure exchange economy analyzed above, we know that the household adapts optimally such that $MRS_{cf} = -\frac{p_c}{p_f}$. On the producers' side, the clothing firm maximizes its profit

$$p_c C - C_C(C), \tag{2.18}$$

where $C_C(\cdot)$ is the clothing firm's cost function (sorry for the double usage of "C" for cost and clothing). Taking the derivative with respect to output C yields the first order condition

$$p_c - C'_C = 0 \tag{2.19}$$

which we can rewrite as

$$MC_{Cloth} = p_c : \tag{2.20}$$

The optimal quantity for a competitive firm is at an output level where its marginal cost equals the output price.

Similarly, profit maximization of the food firm implies

$$MC_{Food} = p_f \tag{2.21}$$

Dividing these two equations through each other and multiplying with -1 therefore implies that

$$-\frac{MC_{Cloth}}{MC_{Food}} = MRT_{cf} = -\frac{p_c}{p_f}.$$

This is exactly the same expression as the $= MRS_{cf}$ of households, so that a market economy achieves an optimal product mix.

2.6 Application: Emissions reduction

Market prices have the very feature that they reflect the underlying scarcity ratios in the economy and help to allocate resources into those of the different uses in which they are most valuable. For example, when there is an excess demand for clothing, the (relative) price of clothing will rise and, as a consequence, additional employment of factors like capital and labor into clothing production becomes more attractive for entrepreneurs.

In this application, we will see how market mechanisms that lead to efficient resource allocation can be used when we want to reduce environmental pollution in a cost efficient way.

Consider the case of SO_2 (sulphur dioxide), one of the main ingredients of "acid rain". SO_2 is produced as an unwanted by-product of many industrial production processes and emitted into the environment. There are however different technologies that allow to filter out some of the SO_2. Some of these technologies are quite cheap, but do not reduce the SO_2 by a lot, and others are very effective, but cost a lot. Moreover, SO_2 is produced in many different places, and some technologies are more efficiently used in some lines of production than in others.

Suppose that we want to reduce the SO_2 pollution by a certain amount The task to find the way to reduce pollution that is (on aggregate) the least costly is quite a complex problem that requires that the social planner (i.e., the government) knows the reduction cost function for each firm.

Suppose that we want to reduce the overall level of pollution that arises from a variety of sources by some fixed amount. Specifically, we assume that there are two firms that emit 1000 tons of SO_2 each. We want to reduce pollution by 200 tons. If firm 1 reduces its emissions by x_1, it incurs a cost of

$$C_1(x_1) = 10x_1 + \frac{x_1^2}{10}. \tag{2.22}$$

Similarly, when firm 2 reduces its emissions by x_2, it incurs a cost of

$$C_2(x_2) = 20x_2 + \frac{x_2^2}{10}. \tag{2.23}$$

We first calculate which reduction allocation minimizes total social cost of pollution reduction. The minimization problem is

$$\min_{x_1,x_2} 10x_1 + \frac{x_1^2}{10} + 20x_2 + \frac{x_2^2}{10} \text{ s.t. } x_1 + x_2 = 200. \tag{2.24}$$

The Lagrange function is

$$10x_1 + \frac{x_1^2}{10} + 20x_2 + \frac{x_2^2}{10} + \lambda[200 - x_1 - x_2]. \tag{2.25}$$

The first order conditions are

$$10 + \frac{x_1}{5} - \lambda = 0 \tag{2.26}$$

$$20 + \frac{x_2}{5} - \lambda = 0 \tag{2.27}$$

Solving both equations for λ and setting them equal gives $10 + \frac{x_1}{5} = 20 + \frac{x_2}{5}$, hence $x_1 = 50 + x_2$. Together with the constraint $x_1 + x_2 = 200$, this yields the solution of

$$x_1 = 125,\ x_2 = 75. \tag{2.28}$$

Hence, firm 1 should reduce its pollution by 125 tons, and firm 2 by 75 tons. The reason why firm 1 should reduce its pollution by more than firm 2 is that the marginal costs of reduction

would be lower in firm 1 than in firm 2, if both firms reduced by the same amount; but such a situation cannot be optimal, since one could decrease x_2 and increase x_1, and so reduce the total cost.

Substituting the solution into the objective function shows that the minimal social cost to reduce pollution by 200 tons is \$ 4875.

For later reference, it is also helpful to note that

$$\lambda = 35. \tag{2.29}$$

The Lagrange multiplier measures the marginal effect of changing the constant in the constraint. Hence, $\lambda = 35$ means that the additional cost that we incur if we tighten the constraint by one unit (i.e., if we increase the reduction amount from 200 to 201) is \$35.

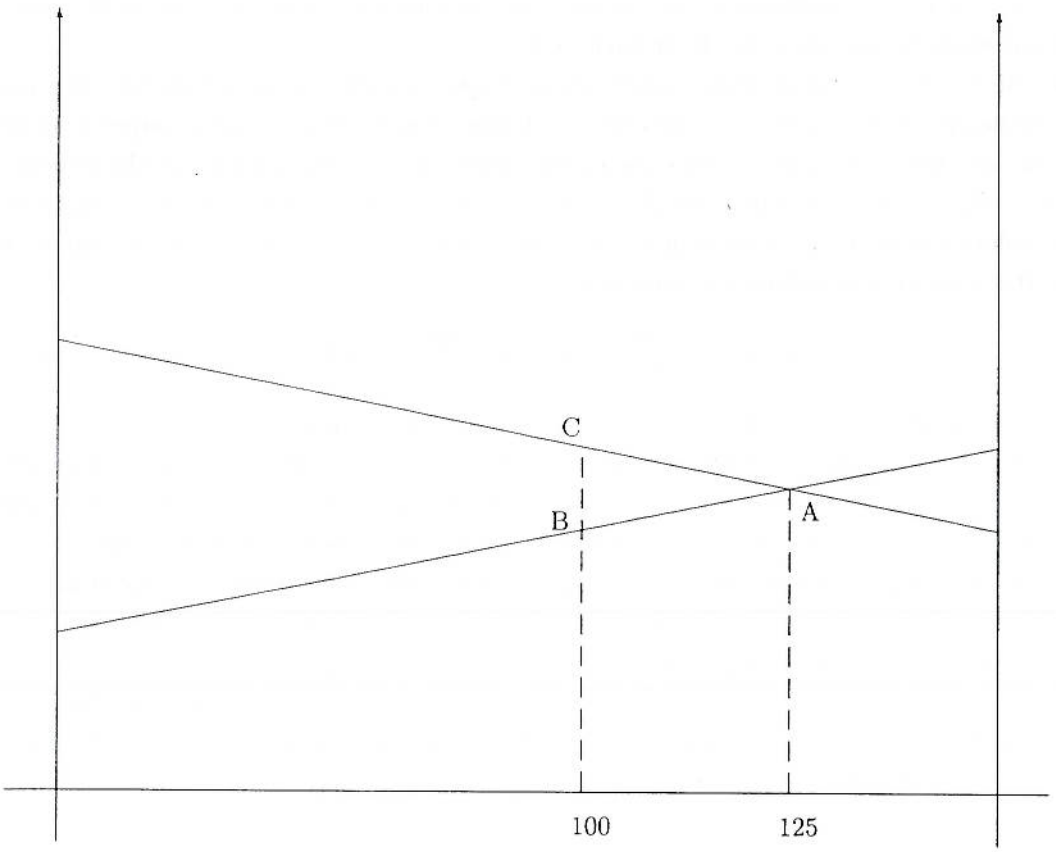


Figure 2.9: Efficient pollution reduction

Figure 2.9 helps to understand the social optimum. The horizontal axis measures the 200 units of pollution that firm 1 and 2 must decrease their pollution in aggregate. The increasing line is the marginal cost of pollution reduction for firm 1, $MC_1 = 10 + \frac{x_1}{5}$. The second firm's marginal cost is $MC_2 = 20 + \frac{x_2}{5}$, and since $x_2 = 200 - x_1$ (by the requirement that both firms together reduce by 200 units), this can be written as $MC_2 = 20 + \frac{200-x_1}{5} = 60 - \frac{x_1}{5}$. This is the decreasing line in Figure 2.9.

The social optimum is located at the point where the two marginal cost curves intersect, at $x_1 = 125$ (and, correspondingly, $x_2 = 75$). Note that, for any allocation of the 200 units of pollution reduction between the two firms (measured by the dividing point between x_1 and the rest of the 200 units), the total cost can be measured as the area below the MC_1 curve up to the dividing point, plus the area below MC_2 from the dividing point on. It is clear that the total area is minimized when the dividing point corresponds to the point where the two marginal cost curves intersect. Any other allocation leads to higher total social costs. For example, if we asked each firm to reduce its pollution by 100 units each, the additional costs (relative to the social optimum) would be measured by the triangle ABC.

We can now turn to some other possible ways to achieve a 200 ton reduction. The first one could be described as a *command-and-control* solution: The state picks some target level for each firm, and the firms have to reduce their pollution by the required amount. In the example, we want to reduce total pollution by 10% from the previous level, and therefore a "natural" control solution is to require each firm to reduce its pollution by 10%, i.e. 100 tons. The total cost of this allocation of pollution reduction is

$$10 \cdot 100 + \frac{100^2}{10} + 20 \cdot 100 + \frac{100^2}{10} = 5000, \tag{2.30}$$

which is of course more than the minimal cost of 4875 calculated above.

Of course, we could in principle also implement the socially optimal solution as a command-and-control solution. However, in practice, this requires that the state has information about the reduction cost functions such that it can calculate the optimal solution. In practice, this extreme amount of knowledge about all different firms is highly unlikely to be available to the state; the following two solutions have the advantage that they rely on decentralized implementation: All that is required is that each firm knows its own reduction cost.

The first solution is called a *Pigou tax*. Suppose that we charge each firm a tax t for each unit of pollution that they emit. When choosing how many units of pollution to avoid, firm 1 then minimizes the cost of reduction minus the tax savings from lower emissions:

$$\min 10x_1 + \frac{x_1^2}{10} - tx_1 \tag{2.31}$$

Taking the derivative yields as first order condition:

$$10 - t + \frac{x_1}{5} = 0, \tag{2.32}$$

hence $x_1 = 5t - 50$. The higher we set t, the more units of pollution will firm 1 reduce. Note however that, if $t < 10$, the firm will not reduce any units, because the lowest marginal cost of doing so (10) is higher than the benefit of doing so, t.

To which amount should we set t? From above, we know that the marginal cost of reduction in the social optimal is \$ 35, and indeed, if we set $t = 35$, we get $x_1 = 125$, just like in the social optimum.

Let us now consider firm 2. It minimizes

$$\min 20x_2 + \frac{x_2^2}{10} - tx_1 \tag{2.33}$$

Taking the derivative yields as first order condition:

$$20 - t + \frac{x_2}{5} = 0, \tag{2.34}$$

hence $x_2 = 5t - 100$. Substituting $t = 35$ yields $x_2 = 75$, again as in the social optimum. Hence, we have shown that, if the state charges a Pigou tax of \$35 per unit of SO_2 emitted, firms will reduce their pollution by 200 tons, and also do this in the most cost-efficient way.

Note that the cost of the Pigou tax for the two firms is substantial. Firm 1 has to pay \$35 for 875 tons, which is \$ 30675. In addition to this, they have to pay abatement costs of $10 \cdot 100 + \frac{100^2}{10} = 2000$. This is much more than firm 1's burden under a command-and-control solution, even if that is inefficient. This is the reason why firms are usually much more in favor of command-and-control solutions to the pollution problem.

A third possible solution is called *tradeable permits*. Under this concept, each firm receives a number of "pollution rights". Each firm needs a permit per ton of SO_2 that it emits, and a firm that wants to pollute more than its initial endowment has to buy the additional permits from the other firm, while a firm that avoids more can sell the permits that it does not need to the other firm.

Suppose, for example, that both firms receive an endowment of 900 permits. Let p be the market price at which permits are traded. If firm 1 reduces its pollution by x_1 units, it can sell $x_1 - 100$ permits; if $x_1 - 100 < 0$, then firm 1 would have to buy so many additional permits.

Firm 1 will maximize its revenue from permits minus its abatement costs:

$$p(x_1 - 100) - 10x_1 - \frac{x_1^2}{10}. \tag{2.35}$$

The first order condition is

$$p - 10 - \frac{x_1}{5} = 0, \tag{2.36}$$

hence

$$x_1 = 5p - 50. \tag{2.37}$$

Similarly, firm 2 maximizes its revenue from permits minus its abatement costs:

$$p(x_2 - 100) - 20x_2 - \frac{x_2^2}{10}. \tag{2.38}$$

The first order condition is

$$p - 20 - \frac{x_2}{5} = 0, \tag{2.39}$$

hence

$$x_2 = 5p - 100. \tag{2.40}$$

In total, the two firms have only 1800 permits, so that they need to avoid 200 tons of SO_2. Therefore,

$$5p - 50 + 5p - 100 = 200. \tag{2.41}$$

Hence, the equilibrium price must be $p = 35$, and thus $x_1 = 125$ and $x_2 = 75$, just as in the social optimum.

2.7 Second theorem of welfare economics

The second theorem of welfare economics states that (under certain conditions) every Pareto optimum can be supported as a market equilibrium with positive prices for all goods. Hence, together with the first theorem of welfare economics, the second theorem shows that there is a one-to-one relation between market equilibria and Pareto optima.

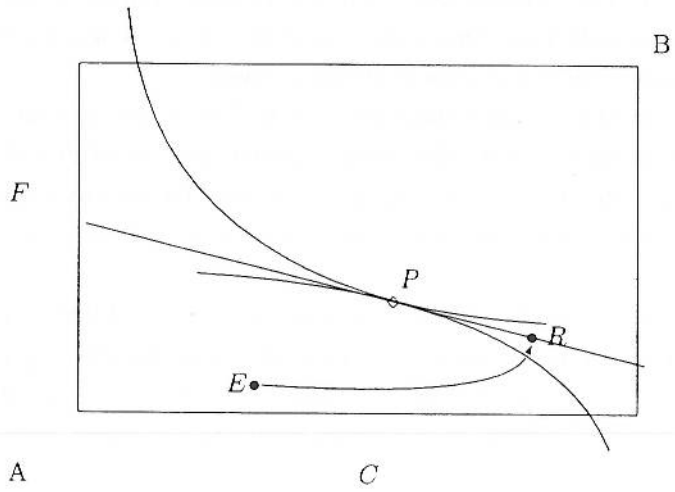


Figure 2.10: Second theorem of welfare economics

In Figure 2.10, the Pareto optimum P can be implemented by redistributing from the initial endowment E to R, and then letting the market operate in which A and B exchange goods so as to move from R to P.

What is the practical implication of the second theorem? Suppose that the government wants to redistribute, because the market outcome would lead to some people being very rich

As a response to a perceived inefficiency of state-owned enterprises, there is often a movement to privatise such firms. Similarly, there is a movement away from cost-plus contracts to fixed price contracts for utilities.

3.6 Cross subsidization and Ramsey pricing

In many cases, the state operates a monopoly as a semi-independent entity. In principle, the state could let the firm run a deficit and cover it from general tax revenue. In practice, politicians are usually reluctant to let the firm accumulate a large deficit and therefore often require that the firm breaks even over all its business lines.

When a firm produces two (or more) different products, then the question arises how we should set the prices of the two products such that we maximize total welfare, subject to the constraint that the firm has to recover its cost. This problem is known as *Ramsey pricing*.

Suppose that the multi-product state firm has the cost function

$$F + c_1 x_1 + c_2 x_2 \tag{3.4}$$

The inverse demand function for good i is denoted by $P_i(x_i)$, which is the inverse of the ordinary demand function $D_i(p_i)$. Our objective is to maximize consumer surplus while covering the fixed cost:

$$\max_x \sum \int_0^{x_i} P_i(s)ds + \lambda[\sum (P_i(x_i) - c_i)x_i - F] \tag{3.5}$$

Differentiating with respect to x_i yields the optimality condition

$$P_i(x_i) + \lambda(P_i - c_i) + \lambda P_i' x_i = 0 \tag{3.6}$$

When we solve for λ and set $i = 1$ and $i = 2$, we get

$$\lambda = \frac{-P_1}{P_1 - c_1 + P_1' x_1} = \frac{-P_2}{P_2 - c_2 + P_2' x_2} \tag{3.7}$$

Inverting both sides gives

$$\frac{P_1 - c_1}{P_1} - \frac{-P_1' x_1}{P_1} = \frac{P_2 - c_2}{P_2} - \frac{-P_2' x_2}{P_2}. \tag{3.8}$$

The second term on both the left and the right hand side is the inverse of the elasticity of demand: $\frac{-P_i' x_i}{P_i} = -\frac{dP_i}{dx_i}\frac{x_i}{P_i} = \frac{1}{\epsilon_i}$. Therefore, we have

$$\frac{P_1 - c_1}{P_1} > \frac{P_2 - c_2}{P_2} \iff \frac{-P_1' x_1}{P_1} > \frac{-P_2' x_2}{P_2}. \tag{3.9}$$

$P_1 - c_1$ is called the *mark-up* on good 1. Dividing by P_1 gives the relative mark-up, which is the profit as a percentage of the price. Equation (3.9) says that the relative mark-up is larger for

good 1 if and only if the elasticity of demand of good 1 is smaller than the elasticity of demand of good 2.

Intuitively, why should the relative mark-up be higher in the less elastic market? In the less elastic market, more revenue can be raised relative to the welfare loss associated with too high prices. (Draw a very steep, hence inelastic, demand curve and consider the effects of a price increase over the marginal cost; the deadweigh loss triangle is small, because the quantity reaction is small in the case of inelastic demand. On the other hand, consider a very flat, that is, very elastic, demand curve. Here, a price increase by the same amount results in a much larger welfare loss triangle, because the quantity reaction is larger. In addition, increasing the price in the elastic market will also raise less revenue than increasing the price in the inelastic market: Suppose that the quantity sold would be the same in both markets if both goods were priced at marginal costs; then the quantity sold in the inelastic market is larger than the quantity in the elastic market (for the same markup).

3.7 Patents

While monopolies are generally bad for social welfare, there is one particular instance in which the state prevents entry into the market, even though competition in the market would be in principle feasible. With a *patent*, an inventor gets the exclusive right to use his invention for some time (usually 14 years in the U.S.).

The reason for this policy is that the monopoly profits provide an incentive to do research. Even if the market for some good operates as a monopoly, the welfare generated is larger than if the good is not invented and therefore the whole market is missing.

Suppose that after the invention is made, the marginal cost with which the good can be produced is constant. If entry in the market is free and everyone can imitate the technology, firms will make zero profit. But this means that the firm who invented the product cannot recover any of its research and development cost for the product. When firms expect that this is the case, no firm has an incentive to spend anything on research and development of new products, and consequently (if these expenditures are necessary for the invention to take place), no new products will be invented without the protection afforded by patents.

Whether or not this rationale for the patent policy is reasonable depends on whether inventions depend on effort or on chance. If research depends very much on effort (in the sense of the conscious decision to spend money on it), then protecting the result by patents is necessary, otherwise no new discoveries will be made. On the other hand, if new inventions sometimes occur without a lot of effort, but rather by chance, then creating a monopoly by issuing a patent is not a good idea.

3.8 Application: Corruption

A nice application of techniques from the analysis of monopoly is to corruption. This section is based on ideas in the paper "Corruption" by Vishny and Shleifer.[1]

Suppose that there is some government activity (like giving a building permit, issuing a passport etc.) for which the demand curve is given by $x = 120 - 2p$ (so that the marginal willingness to pay is $60 - \frac{1}{2}x$). While there is no cost of production, the access to the service is controlled by some government official(s) who set a "price" (i.e. bribe) for the service.

Suppose first that there is a strictly organized system of corruption in which one government official sets a price. The official then maximizes

$$\max_p (120 - 2p)p. \tag{3.10}$$

Taking the derivative yields

$$120 - 4p = 0 \Rightarrow p = 30 \tag{3.11}$$

Hence, the price that the government official charges is \$30, and for this price $x = 120 - 2 \cdot 30 = 60$ units are sold. The total welfare generated in this market is

$$\int_0^{60} [60 - \frac{1}{2}x]dx = [60x - \frac{1}{4}x^2]_{x=0}^{x=60} = 3600 - 900 = 2700. \tag{3.12}$$

Consider now what happens if there are two government officials who are necessary to receive the service and who operate independently. For example, suppose that in order to operate a shop, a police permit and a tax authority identification number are necessary, and both are sold separately. Note that there is no point in just buying one of these items, the buyer needs both in order to be able to operate; consequently, the buyer considers the *sum of the two prices* and compares it to his willingness to pay: The demand function then is $x = 120 - 2(p_1 + p_2)$.

The first government official (say, the police) maximizes

$$\max_{p_1} (120 - 2p_1 - 2p_2)p_1 \tag{3.13}$$

which yields the following first order condition:

$$120 - 4p_1 - 2p_2 = 0 \tag{3.14}$$

Similarly, the second government official (say, the tax authority) maximizes

$$\max_{p_2} (120 - 2p_1 - 2p_2)p_2 \tag{3.15}$$

which yields the following first order condition:

$$120 - 2p_1 - 4p_2 = 0. \tag{3.16}$$

[1] A. Shleifer and R. Vishny, 1993, "Corruption", QJE, 108 (3), 599-617.

Solving (3.14) and (3.16) simultaneously for p_1 and p_2 yields $p_1 = p_2 = 20$. Hence, the total price that a citizen has to pay in order to operate a shop is $p_1 + p_2 = 40$ and hence higher than before. Consequently, the quantity sold decreases to $x = 120 - 2 \cdot 40 = 40$, and total welfare generated in this market is

$$\int_0^{40} [60 - \frac{1}{2}x]dx = [60x - \frac{1}{4}x^2]_{x=0}^{x=40} = 2400 - 400 = 2000. \tag{3.17}$$

Note that an "strictly organized" corrupt society in which only one official sets a price for the government service generates a higher welfare for society than a system in which several agents all set prices independently. The reason is that each of these independent agents exerts a negative externality on the other one when he increases his own price, because a price increase by agent 1 does not only decrease the demand for permits that agent 1 faces, but also the demand for agent 2's permits (and vice versa). When deciding how to set his price, agent 1 ignores his negative effect on agent 2's business and therefore sets his price too high *even if one only considers the profit of the bribed bureaucrats.* Of course, from the consumers' point of view, a single monopolistic bureaucrat is far from optimal, but still, a unitary bureaucratic bribing system is much better than several independent agencies.

3.9 Introduction to game theory

In this section, we want to give a very brief introduction to game theory. Game theory is a set of tools used to analyze problems of strategic interaction between several players. In economics, there are very many realistic situations that can be interpreted as "games".

Games are classified according to whether they are static or dynamic. In static games, all players move at the same time, while in dynamic games at least some players move after having observed the previous moves of other players. Another dimension of classifications is whether a game has complete or incomplete information. In complete information games, all players know each others' objective functions and the actions available to each player, while this is not the case in incomplete information games. We will deal here only with static games of complete information, the simplest type of games.

The Prisoners' Dilemma. A famous type of game is the so-called *Prisoner's dilemma.* In the first description of the game, two bank robbers are caught by the police, but there is not enough evidence to convict them of bank robbery, if both remain quiet. For this, the police needs (at least) one of the two to confess. However, the police have enough evidence to convict both prisoners of illegal weapons possession (which carries a relatively low penalty).

If both prisoners remain quiet, they receive a utility of 2 each. The police makes the following offer to each prisoner: *"If you confess while the other robber does not, you will go free (giving you a utility of 3); your colleague however will receive a very long sentence in this case, because*

public good.

In order to find a Pareto optimum, we maximize a weighted sum of the two individuals' utilities:[1]

$$\max_{g_1,g_2} a_1 U_1(g_1 + g_2, w_1 - g_1) + a_2 U_2(g_1 + g_2, w_2 - g_2) \tag{4.2}$$

The first order conditions are

$$a_1 \frac{\partial U_1}{\partial G} - a_1 \frac{\partial U_1}{\partial x_1} + a_2 \frac{\partial U_2}{\partial G} = 0 \tag{4.3}$$

$$a_1 \frac{\partial U_1}{\partial G} + a_2 \frac{\partial U_2}{\partial G} - a_2 \frac{\partial U_2}{\partial x_2} = 0 \tag{4.4}$$

We can rewrite the first order conditions as follows:

$$a_1 \frac{\partial U_1}{\partial G} + a_2 \frac{\partial U_2}{\partial G} = a_1 \frac{\partial U_1}{\partial x_1} \tag{4.5}$$

$$a_1 \frac{\partial U_1}{\partial G} + a_2 \frac{\partial U_2}{\partial G} = a_2 \frac{\partial U_2}{\partial x_2} \tag{4.6}$$

Since the left hand sides are the same, the right hand sides must be equal, too. Take the first equation, and divide both sides by $a_1 \frac{\partial U_1}{\partial x_1}$ (on the left hand side, divide the first term by $a_1 \frac{\partial U_1}{\partial x_1}$, and the second term by $a_2 \frac{\partial U_2}{\partial x_2}$ which is the same, as argued above). This yields

$$\frac{\frac{\partial U_1}{\partial G}}{\frac{\partial U_1}{\partial x_1}} + \frac{\frac{\partial U_2}{\partial G}}{\frac{\partial U_2}{\partial x_2}} = 1 \tag{4.7}$$

The first term on the left hand side is (the absolute value of) the marginal rate of substitution dx/dG of individual 1 between public and private good. It tells us how many units of the private good individual 1 is willing to give up for one more unit of the public good. The second term on the left hand side is the same expression for individual 2. The optimality condition then says that the sum of the marginal rates of substitution needs to be equal to 1, which is the marginal cost of the public good. Since we can consider the marginal rate of substitution as a marginal willingness to pay, in terms of the other good, this result has essentially the same interpretation as the marginal benefit expression above.

4.3 Private provision of public goods

Let us now consider what happens if both players decide individually with their own private interest in mind whether and how much to contribute to the public good.

As the first simple example, consider the following setup: All n players have two feasible actions, to "contribute" or "not to contribute". If a player contributes, he has to pay \$100,

[1] Alternatively, we could also maximize the utility of individual 1 subject to the constraint that individual 2 needs to reach a particular utility level. This problem has a Lagrangean that is (up to a constant) the same as the weighted sum of utilities used below, and therefore leads to the same result.

but every player including the contributor himself, there is a \$80 benefit. If several players contribute, then each player receives \$80 times the number of contributors, minus his cost (if applicable). Evidently, as long as there are at least two players, it would be very beneficial if all players "contribute". However, the payoff structure of the game is the same as in the Prisoners' Dilemma in Section 3.9. It is a dominant strategy for each player not to contribute.

Consider player 1, and suppose that, from the other $n-1$ players, k contribute. If player 1 does not contribute, he receives $80k$. If he contributes, he receives $80(k+1)-100$, because there are now $k+1$ contributors, but player 1 has to pay his contribution cost; player 1's payoff can be simplified to $80k-20 < 80k$, so whatever the number of other people who contribute, the optimal action for player 1 is not to contribute. Of course, the same argument holds for all other players, and thus no player will contribute in equilibrium.

For two players, we can capture the payoffs in the Table 4.1. It is easy to check that (don't contribute, don't contribute) is the unique equilibrium of this game.

		Player 2	
		contribute	don't contr.
Player 1	contribute	(60,60)	(-20,80)
	don't contr.	(80,-20)	(0,0)

Table 4.1: Payoffs in the public good game

As our second example, consider the case of a continuous public good. Suppose that A has a marginal benefit from the public good given by $MB_A = 10 - G$; B has a marginal benefit given by $MB_B = 8 - G$. The marginal cost of providing one unit of the public good is 4.

To determine the efficient quantity, we simply add the two marginal benefits to get $MB_A + MB_B = 18 - 2G$ and set this equal to the marginal cost of 4. Hence, the efficient quantity is $G^* = 7$.

A Nash equilibrium of a game is a pair of actions (one for each player) that are mutually best responses, i.e., given the actions played by the other players, no player could increase his payoff by choosing a different action. What is the Nash equilibrium of the public good provision game when both players decide simultaneously how much to contribute?

We claim that "A provides 6 units of the public good and B provides 0" is a Nash equilibrium. To check this claim, let us suppose that B does not contribute anything, and analyze what the best action for A is. The best A can do in this situation is to buy G such that his own marginal benefit is equal to the marginal cost, hence to choose $G = 6$ as contribution. To see this formally, note that, if A's marginal benefit is $10 - G$, then his benefit is given by $10G - 0.5G^2 = 10(g_1 + g_2) - 0.5(g_1 + g_2)^2$. When A maximizes benefit minus cost, the first order optimality condition is $10 - (g_1 + g_2) - 4 = 0$, or marginal benefit equal to marginal cost. Hence,

given that B does not contribute anything, it is actually optimal for A to contribute 6 units.

Second, given that A contributes 6 units, does B have an incentive to behave as claimed and not contribute anything? B's gross benefit from the public good given that G units are provided is given by $8G - 0.GX^2$. If A contributes 6 units and B contributes g_B units, B's net utility is

$$8(6 + g_B) - 0.5(6 + g_B)^2 - 4g_B \tag{4.8}$$

Differentiating with respect to g_B yields

$$8 - (6 + g_B) - 4 = -2 - g_B. \tag{4.9}$$

This means that, even starting from $g_B = 0$, the marginal net benefit of making a contribution for B is negative. Another way to see this is to note that, at $G = 6$, B's marginal gross benefit from the public good is only 2, and therefore it is not worth it to spend the cost of 4 to even contribute the first additional unit.

This is in fact the unique Nash equilibrium. To see this, note that there is never any level of the public good such that both A's and B's marginal gross benefit is equal to 4 (the marginal cost). But this would be a necessary condition for both players to be willing to contribute positive amounts. Consequently, in the Nash equilibrium, the player with the higher marginal benefit pays *everything*, the other player just benefits and pays nothing. Note also that the public good is under-provided, in the sense that the equilibrium quantity of the public good is smaller than the efficient quantity.

4.4 Clarke–Groves mechanism

In the last section, we have seen that the amount of a public good that is provided through private voluntary contributions is likely to be too low from a social point of view. If the state knows the marginal benefit functions of people, it can intervene and just provide the efficient quantity of public goods. However, in practice, we face the following problem: How could the state find out how much of the public good to supply, if individual demand functions are unobservable for outsiders? (Of course, all people know their own utility.)

To make the issue more concrete, consider the following problem. There is an indivisible public good, and if the good is provided, its cost is 1. The public good (if provided) generates a benefit $v_A \in (0, 1)$ for A and $v_B \in (0, 1)$ for B. Only A knows v_A and only B knows v_B. We call v_A and v_B the players' *types*.

Clearly, the efficient solution is to provide the good if and only if $v_A + v_B > 1$, but in order to know whether this inequality is satisfied, we need to ask A and B for their respective benefits, and we have to provide them with an incentive to reveal their benefit truthfully.

To see that there might be a problem, namely that people misrepresent their preferences, consider the following mechanism.

1. Both people are asked about their type. They report values of m_A and m_B (m stands for message), respectively. Players may choose to report their true type, or may report some other value. (Since the players themselves are the only persons to know their types, there is no way how we could force them to tell the truth.)

2. If $m_A + m_B > 1$, the good is provided; A pays $\frac{m_A}{m_A+m_B}$, B pays $\frac{m_B}{m_A+m_B}$

If both people tell the truth, this mechanism implements the social optimum. Also, the payments required by the mechanism appear basically "fair", because the cost share of each player is proportional to his share of the benefits. However, will people tell the truth under this mechanism?

Consider A with type v_A, and suppose that B tells the truth. If A reports to be of type m, A's expected utility is

$$\int_{1-m}^{1} \left[v_A - \frac{m}{m+v_B} \right] f(v_B) dv_B,$$

where $f(v_B)$ is the density function of the distribution of B's possible types. If A reports m, then the public good is provided if and only if $v_B > 1-m$ (because otherwise the sum of the two reports would be smaller than 1). In case the public good is provided, A's payment is $\frac{m}{m+v_B}$, so that the term in square brackets is A's surplus. Multiplying with the probability of each type of player B and summing up[2] yields the above expression.

To find A's optimal report, take the derivative with respect to m. This yields[3]

$$\left[v_A - \frac{m}{m+1-m} \right] f(1-m) - \int_{1-m}^{1} \frac{v_B}{(m+v_B)^2} f(v_B) dv_B.$$

Evaluated at $m = v_A$, the first term is zero, and hence the derivative is negative. It follows that it is better to set $m < v_A$.

Intuitively, suppose A sets $m = v_A$, but considers a small "lie" understating his true valuation. Most likely, this will not change whether the public good is provided or not. If the public good is provided even after A's small lie, then the only effect of the lie is that it reduces A's payment. This is a sizable effect (the second term). There is also some chance that the public good would have been provided if A had told the truth about his valuation, but is not provided if A lies. In principle, this provides some deterrent for A to understate his valuation, because A might lose a public good that he likes. However, suppose that $m \approx v_A$; then, A's payment if he tells the truth (in those cases where the public good is not provided when A lies) is large (namely approximately v_A), and thus, A would not have a significant surplus even if the public good is provided. Hence, the expected size of the loss associated with this contingency is very small.

[2] Integrating is essentially the same as summing up.

[3] For taking the derivative with respect to the m that stands in the lower limit of integration, note that Leibnitz' rule states that $\frac{\partial}{\partial a} \int_a^b f(x)dx = -f(a)$ and $\frac{\partial}{\partial b} \int_a^b f(x)dx = f(b)$.

Note that we have not actually calculated what *will* happen in equilibrium, since we have not derived an equilibrium; for our purposes, it is sufficient to know that we cannot implement the efficient solution with this mechanism.

Let us now consider a different mechanism in this situation, called the *Clarke-Groves mechanism*, which works as follows:

1. Both people announce m_A and m_B as their willingness to pay (they can, of course, lie, just as before).

2. If $m_A + m_B > 1$, the good is provided, A pays $(1 - m_B)$ and B pays $(1 - m_A)$

Observe first that the report m_A affects A's payoff only if it changes whether the good is provided; the price A has to pay (if the good is provided) is independent of m_A and depends only on B's report. It is this property that makes the difference to the mechanism above.

Suppose A knew B's report m_B. If $v_A + m_B > 1$, then announcing $m_A = v_A$ is optimal for A: If A chooses $m_A = v_A$, then the public good will be provided and A's surplus is $v_A - (1 - m_B) > 0$. Alternatively, A lies and chooses m_A such that $m_A + m_B < 1$, then the public good will not be provided and A will receive a utility of 0 (which is worse for him than what he gets if he tells the truth). Of course, A could also misrepresent his type, but in a way that $m_A + m_B > 1$: In this case, A's payoff is the same as if he tells the truth, since neither the public good provision nor the amount that A has to pay changes. The main point is, however, that A cannot strictly gain by misrepresenting his preferences.

Second, suppose that $v_A + m_B < 1$. In this case, announcing $m_A = v_A$ is again optimal for A: If A announces $m_A = v_A$, then the public good will not be provided, so that A receives a utility of 0. Alternatively, A could claim to have m_A such that $m_A + m_B > 1$ and the public good is provided.[4] However, this means that A has to pay a price $1 - m_B$ for a benefit of only v_A from the public good, and since $v_A - (1 - m_B) < 0$, this is strictly worse for A than announcing his true type.

Up to now, we have assumed that A knows B's announcement m_B, which is, of course, not realistic. However, we have shown that *independent of what m_B actually is*, it is optimal for A to announce his true type. Therefore, it does not matter that, in reality, A does not know the announcement of B. Note also that it does not matter for A whether B reveals his type truthfully or lies, so whatever B does, A's optimal action is to reveal his true type.

Of course, a symmetric argument implies that also B will reveal his true type in the Clarke-Groves mechanism. Since both players announce the truth, and the public good is provided if and only if this is efficient if the announcements are truthful, the efficient solution can be implemented with the Clarke-Groves mechanism. It is remarkable that this is true no matter what the distribution of v_A and v_B is, so this result holds in a very general setup (which is

[4] Evidently, misrepresenting his preferences in a way that $m_A < 1 - m_B$ does not change the outcome at all and therefore does not help A, relative to what he gets when he reveals his type truthfully.

important since in reality, it may also not be too easy to specify what a realistic distribution for the v_i values is).

Intuitively, the Clarke-Groves mechanism makes each individual the "residual claimant" for the social surplus. Assuming that the other individual (say, B) tells the truth, the "social surplus excluding A" is $v_B - 1$ (that is, B's payoff minus the cost of provision, 1). In the Clarke-Groves mechanism, A has the option to pay the negative of this "social surplus excluding A" $(1 - v_B)$ as a price to receive the public good. Faced with this decision, A wants to get the public good if and only if his surplus v_A is larger than this price. Announcing his true willingness to pay, v_A, is a way to secure that the public good is provided if and only if $v_A > 1 - v_B$.

Note that the Clarke-Groves mechanism does not have a *balanced budget* in the following sense: If $v_A + v_B > 1$ so that the public good is provided, then both individuals pay $(1 - v_B) + (1 - v_A) = 2 - v_A - v_B < 1$, hence less than the cost of the public good. Therefore, in order to be able to pay for the public good, a third party ("state") has to put in some money. The state could charge from both people an additional lump sum payment (i.e., the same amount, whether or not the good is provided) to offset this. However, it is not possible to construct a mechanism that gives incentives for truthful revelation *and* has a balanced budget for all possible realizations of the types.

4.5 Applications

4.5.1 Private provision of public goods: Open source software

(see papers posted on class webpage)

4.5.2 Importance of public goods for human history: "Guns, germs and steel"

In the provocative book "Guns, germs and steel", Jared Diamond provides an analysis of human history starting from the development of agriculture to today. In particular, he asks how Eurasians and their descendants came to dominate the world after the 15th century.

In the 13,000 years since the end of the last Ice Age, some parts of the world developed literate industrial societies with metal tools other parts developed only non-literate farming societies and still others retained societies of hunter-gatherers with stone tools. Those historical inequalities have cast long shadows on the modern world, because the literate societies with metal tools have conquered or exterminated the other societies.

Yali, a New Guinea politician asked Diamond "Why is it that you white people developed so much manufactured goods and brought it to New Guinea, but we had little goods of our own?" To rephrase, "why did wealth and power become distributed as they now are, rather than in some other way? For instance, why weren't Native Americans, Africans, and Aboriginal

As can be seen from Figure 5.5, the equilibrium level of research x' is lower than the socially optimal level x^*, since the firm does not take the positive externalities into account.

The ways to deal with positive externalities very much mirror the possibilities to deal with negative externalities. While a Pigou tax is used to reduce the amount of negative externalities, the state can subsidize the positive externality generating activity. In fact, this is one of the reasons why the state often subsidizes private companies doing research, and also why the state subsidizes research in universities.

In terms of private solution, a merger between the firms involved again provides the correct incentives (but, again, the applicability of this solution may be limited; see the discussion in the last section). A way to define property rights in this application would be the patent system; however, as argued above, not everything can be patented (and also, it is not clear that every new idea should be patented even if this were possible).

5.5 Resources with non-excludable access: The commons

In the middle ages, the *"commons"* were a meadow which belonged to all farmers of a community together and everyone could choose to let livestock graze on the commons. In a way, this served as a social insurance net for farmers who were too poor to afford their own meadows. Today, we refer to resources as "commons" if the access is non-excludable, but rival. Modern day examples for commons include fishing in the world's oceans and the usage of streets in cities during the rush hour.

To analyze these problems, let us consider the following example, which looks at the original common meadow problem. Suppose that the price of a cow is 5. Cows produce milk, which has a price normalized to 1. Let x_i denote the number of cows that farmer i chooses to graze on the commons, and let $X = \sum_{i=1}^{n} x_i$ be the total number of cows.

Each cow produces $20 - \frac{1}{10}X$ units of milk. The reason why the milk output of a given cow is decreasing in the total number of cows is of course that, the more cows there are, the less grass per cow is available, and insufficiently fed cows produce less milk. (The linear functional form is, of course, only taken to simplify the analysis.)

Let us first find the cooperative solution that would maximize the joint profit of all village farmers. This optimization problem is

$$\max_X [20 - \frac{1}{10}X - 5]X \tag{5.1}$$

Taking the derivative and setting it equal to zero gives

$$15 - \frac{2}{10}X = 0 \Rightarrow X = 75. \tag{5.2}$$

Hence, 75 cows should graze on the commons, for a total profit of $75 \cdot 7.5 = 562.50$.

n	1	2	4	9	∞
X	75	100	120	135	150
PPC	7.5	5	3	1.5	0
TP	562.5	500	360	202.5	0

Table 5.1: Usage X, profit per cow and total profit with n farmers

Let us now consider what happens when the n farmers decide simultaneously how many cows to graze on the commons. Given the other farmers' decisions, farmer i maximizes his profit:

$$\max_{x_i}[20 - \frac{1}{10}(x_1 + x_2 + \cdots + x_i + \cdots + x_n) - 5]x_i \tag{5.3}$$

We find the condition for an optimum by differentiating with respect to x_i:

$$15 - \frac{1}{10}(x_1 + x_2 + \cdots + 2x_i + \cdots + x_n) = 0 \tag{5.4}$$

Since we have n conditions (one for each farmer), we have in principle an $n \times n$ linear equation system. The easiest way to solve such a system, given that all of these equations look alike, is to invoke symmetry: In an equilibrium, it is plausible that every farmer has the same number x of cows on the meadow. We can then substitute $x_j = x$ for all j in equation (5.4); note that there are n terms in the bracket, and that there is a factor of 2 before x_i, so that we get $(n+1)x$ for the term in brackets. Overall, we get

$$15 - \frac{1}{10}(n+1)x = 0 \Rightarrow x = \frac{150}{n+1}. \tag{5.5}$$

Therefore, when the number of independent farmers is n, the total number of cows is $150\frac{n}{n+1}$. Table 5.1 reports the total number of cows, the profit per cow and the total profit of all farmers as a function of n, the number of farmers.

If there is only one farmer, this farmer chooses the number of his cows to be equal to the social optimum. The reason is that he receives all social benefits and pays all social costs of his decisions. The more farmers there are, the higher is the number of cows that graze on the commons. The reason is that each farmer imposes a negative externality on the other farmers. When there are few farmers around (say, $n = 2$), then a large percentage of the negative effects of adding an extra cow (namely that all other cows give less milk) still hits the farmer himself. In the case of $n = 2$, 50% of the cows belong to a farmer, and so only the remaining 50% are "externalized". If, instead $n = 9$, then 8/9 of the negative effect of lower milk production per cow affect the cows of other farmers and are therefore disregarded when deciding on the quantity.

In the limit of very many farmers, $X = 150$, and the total profit is zero! The reason is that *all* negative effects now hit other people, and therefore a farmer will choose to add cows as long as the milk yield from this cow is sufficient to pay for the cost of a cow. In this case, common access resources suffer from extreme over-utilization and all the potential benefits from the commons are lost.

The proof of this theorem is quite immediate. Suppose, for example, that there is an election over the median voter's favorite candidate and some other candidate who is to the right of that candidate. All voters whose bliss points are located to the left of the median's bliss point prefer the median's candidate over the right wing candidate, and since they are (together with the median) one more than half of the electorate, they will win the election over any right wing candidate. A symmetric argument shows that a left wing candidate cannot beat the median's favorite candidate.

7.4.2 Example: Voting on public good provision

Consider the following example: There are N voters, everyone has the same utility function $U(x, G) = x + \ln(G)$, where x is the amount of a private consumption good, and G is the amount of the public good provided. Individual i has a gross income of y_i which is subject to proportional taxation at rate τ. The government uses the tax revenues to buy an amount G of the public good, so that $G = \sum \tau y_j$.

We are interested in the preferred tax rates of individuals. The indirect utility of individual i as a function of the tax rate τ is

$$V_i(\tau) = (1 - \tau)y_i + \ln(\tau \sum y_j) \tag{7.4}$$

In order to find the optimal tax rate for individual i, we take the derivative of (7.4) to get

$$-y_i + \frac{1}{\tau} = 0 \Rightarrow \tau_i^* = \frac{1}{y_i} \tag{7.5}$$

This shows that richer voters prefer a smaller tax rate. Moreover, differentiating the derivative a second time gives $-1/\tau^2 < 0$ so that i's indirect utility is globally concave in τ and therefore really has a single peak at $\tau_i^* = \frac{1}{y_i}$.

If this society determines its tax rate by voting, then we expect that the median rich individual determines the equilibrium tax rate, so that we get $\tau = 1/y_m$.

It is interesting to compare this tax rate with the social optimum, in which the social planner maximizes the sum of all citizens' utility:

$$\max[\sum_{i=1}^{N}(1 - \tau)y_i] + N \ln(\tau \sum y_j) \tag{7.6}$$

which gives the following first order condition:

$$-Y + N\frac{1}{\tau} = 0, \tag{7.7}$$

where $Y \equiv \sum y_i$ is the total income of all citizens. Hence, the socially optimal tax rate is $\tau_{SO} = \frac{1}{y}$, where $y = Y/N$ is the average income.

How does this tax rate compare to the Condorcet winner? Empirically, in (almost) all societies, the average income is higher than the median income. The reason is that there are a number of very rich individuals who contribute a lot to the average income, but hardly at all change the median income. If we take $y_m < y$ as given, then $\tau > \tau_{SO}$ and $G > G_{SO}$, so that a democratic society will tend to provide too much of the public good in this example.

The intuitive reason is that the median voter gets a share of $1/N$ of the total benefit from the public good (since all individuals receive the same benefit), but pays less than $1/N$ of the total taxes, because his median income is less than the average income. This provides an incentive to overspend for the median voter.

Note that this simple model assumes proportional taxation, i.e. each individual pays the same tax rate (this would arise, for example, if sales taxes were the only tax to finance public good provision by the government. With *progressive taxation*, i.e. a system in which rich people pay a higher *percentage* of their income as taxes than poor people, the overspending result is even stronger.

7.4.3 Candidate competition

A particularly important application of the median voter theorem is to competition for political office between two candidates. Suppose that politicians can make promises to voters as to which policy they will implement if they are elected to office. In equilibrium, both politicians will try to get as close to the median voter as possible, because the median voter is decisive for which candidate wins. Note that the reason for this is not that there are necessarily many people with moderate policy preferences. The argument would also hold if there are actually very few moderates: Even in this case, there is no point in catering to more extreme preferences; left-wing extremists, say, will vote anyway for the politician who is (slightly) more left than the other candidate. The important votes to win are in the mid of the political spectrum.

7.5 Multidimensionality and the median voter theorem:

The median voter theorem is often interpreted in the sense that for "realistic" preferences, the potential problems indicated by Arrow's impossibility theorem do not exist. For single peaked preferences, a Condorcet winner exists. Unfortunately, the median voter theorem does not carry over to the case that there are several dimensions of policy, even if preferences in this more complex policy space are still single-peaked.

As an example of a multidimensional policy space, consider the following example. Dimension 1 is the tax rate, as in our previous example. However, there is also a second dimension that we can interpret as social issues (say, abortion rights) and voters differ in their preferences over how conservative or liberal the politician should be in this dimension.

In an equilibrium, it must not be worthwhile for D-types to imitate S-types; otherwise, all D-types would in fact imitate and then the belief of employers that someone who has obtained education $\bar{x}$ is not justified. This implies that in a separating equilibrium, we must have $\bar{x} \geq 10/\delta$. The lowest education level that can support a separating equilibrium is therefore $\bar{x} = 10/\delta$.

The other question in a separating equilibrium is whether S-types do in fact find it optimal to choose $x = \bar{x}$? To check this, note that an S-type gets $20 - 10/\delta$. Since $\delta > 1$, this is more than 10, which is an S-type's payoff if he imitates a D-type and does not get any education.

From a social point of view, education is wasteful in this model. Hence, if we weigh all individuals' utility the same, the state should prohibit education (this, of course depends on the assumption that education is not productive at all). As for the different types, note first that D-types always prefer the pooling equilibrium, because they just get a higher wage in the pooling than in the separating equilibrium. For S-types, the situation is a bit more complicated: S-types prefer the pooling equilibrium if $10(1 + p) > 20 - 10/\delta$, otherwise, they prefer the separating equilibrium.

6.4 Moral Hazard

6.4.1 A principal-agent model of moral hazard

Economists talk about *moral hazard* in situations in which, after the contract is signed, the behavior of one party is affected by the terms of the contract, and the change in behavior affects the other party's utility. The term "moral hazard" comes from the insurance sector and refers to changes in loss prevention activities. Once an individual has bought insurance, he may not be quite as careful any more in order to prevent the loss as he would be without insurance. Insurance companies consider this behavior as "immoral", hence the term.

One of the most important applications of moral hazard models is to incentives for workers in firms. The owner of the firm, whom we call the "Principal" hires an agent to work for him; think of a manager who actually directs the day-to-day operations of the firm. For simplicity, let us assume that there are two possible output y_h and y_l, with $y_h > y_l$; we denote the probability of a high output with p. The output level is observable and verifiable and so the salary can depend on the level of output; let s_h denote the worker's salary when the output is y_h, and s_l when the output is y_l.

The agent can choose between two possible effort levels, e_h and e_l, with $e_h > e_l$. High effort increases the probability of success, but is also more costly to the worker. Specifically, the worker's utility is

$$u = p_h\sqrt{s_h} + p_l\sqrt{s_l} - c_e, \tag{6.1}$$

where c_e is the cost of effort (either c_l when effort is low, or c_h when effort is high). The first two terms are the worker's expected utility from wage payments. We assume that the worker is *risk averse*, and has a utility function equal to the square root function.

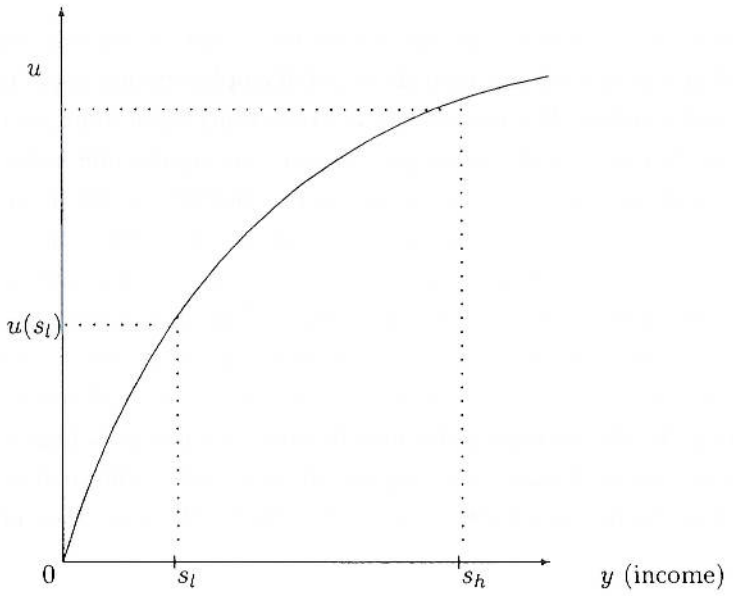


Figure 6.1: Expected utility

Why do we assume this particular utility function for the worker, and what are the implications? The expected utility of the worker is a weighted average between $\sqrt{s_l}$ and $\sqrt{s_h}$, where the weights are the respective probabilities for the events. Graphically, the expected utility corresponds to a point on the line connecting the two points $(s_l, \sqrt{s_l})$ and $(s_h, \sqrt{s_h})$, namely $(p_l s_l + p_h s_h, p_l\sqrt{s_l} + p_h\sqrt{s_h})$. The first coordinate of this point is the expected income, and the second coordinate is expected utility. With the square root function (and other functions called *concave*; those having decreasing first derivatives), the connecting line between two points that are on the function is always below the function. That means that expected utility is below the value of the utility function, evaluated at expected income: The worker would strictly prefer to receive a certain amount of income that is equal to the expected income, to a "lottery over income" (in which he sometimes gets more money and sometimes less, but which has the same expected value). We call such a preference for certainty "risk aversion", and the square root function for expected utility is one function that has this property.

The principal, on the other hand, has a linear utility function, hence does not mind whether he has a lottery or a certain amount of income (with the same expected value). It is generally accepted that risk aversion is decreasing in income, i.e., rich people are (for a given risk) less risk averse than poor people. Thus, the assumption that the principal is less risk-averse than the agent is quite plausible (even though the assumption that he is completely risk-neutral is, of course, less so; but that is not important). From the point of view of efficient risk-sharing, the principal should take all the risk away from the agent. We can also say that he should "insure" the agent against income risk.

Returning to the moral hazard problem, the relation between effort and success is given by the following Table 6.1.

	$y = y_h$	$y = y_l$
$e = e_h$	0.7	0.3
$e = e_l$	0.4	0.6

Table 6.1: Relation between effort and output

By exerting high effort, the agent is able to increase the success probability from 0.4 to 0.7. Note that high effort increases success probability, but failure remains a possibility. Similarly, a shirking worker may simply be lucky.

We now analyze how the optimal contract looks like. There are several cases to look at.

Case 1: Effort is observable and verifiable. In this case, the contract can specify the effort level that the agent is required to choose. If, say, the contract specifies that the agent has to exert high effort, but chooses only low effort, then the principal can take him to court and receive a large penalty payment. hence, without loss of generality, we can assume that, if the agent signs a contract that requires him to exert him effort, then he will exert high effort in equilibrium.[3] In addition, the contract may also specify that the wage depends on the output level that has been realized.

Suppose first that the contract specifies that the agent has to exert high effort. The optimal contract for the principal chooses the wage level s_h and s_l for high and low output, so as to maximize

$$\max 0.7(y_h - s_h) + 0.3(y_l - s_l) \tag{6.2}$$

subject to the constraint that the worker is willing to sign the contract:

$$0.7\sqrt{s_h} + 0.3\sqrt{s_l} - c_h \geq \bar{u}. \tag{6.3}$$

Here, $\bar{u}$ is the *outside option* of the worker, that is, the utility that the worker can get if he works for another firm. Differentiating with respect to s_h and s_l yields the following first order conditions:

$$-0.7 + 0.7\frac{\lambda}{2\sqrt{s_h}} = 0 \tag{6.4}$$

$$-0.3 + 0.3\frac{\lambda}{2\sqrt{s_l}} = 0 \tag{6.5}$$

[3] Of course it is not guaranteed that the optimal contract will indeed require high effort. Since the agent has a higher cost of effort when exerting high effort, the wage payment necessary to induce the agent to sign a high effort contract is likely to be larger than the wage payment under a low effort contract.

Solving both equations for λ yields $\lambda = 2\sqrt{s_h} = 2\sqrt{s_l}$ which implies $s_h = s_l = \lambda^2/4$. We now take this and substitute into the constraint (6.3) to get $\lambda/2 = \bar{u} + c_h \Rightarrow s_h = s_l = (\bar{u} + c_h)^2$.

Let us now consider the optimal salaries under low and high output if the contract specifies low effort. The principal's problem is

$$\max 0.4(y_h - s_h) + 0.6(y_l - s_l), \tag{6.6}$$

subject to the participation constraint

$$0.4\sqrt{s_h} + 0.6\sqrt{s_l} - c_l \geq \bar{u}. \tag{6.7}$$

The first order conditions are

$$-0.4 + 0.4\frac{\lambda}{2\sqrt{s_h}} = 0 \tag{6.8}$$

$$-0.6 + 0.6\frac{\lambda}{2\sqrt{s_l}} = 0, \tag{6.9}$$

From which we again find that $s_h = s_l = \lambda^2/4$. When we substitute this into the participation constraint, we get $\lambda/2 = \bar{u} + c_l \Rightarrow s_h = s_l = (\bar{u} + c_l)^2$

Note that both contracts have "full insurance", that is, the agent receives the same wage in both output states; the principal assumes all the risk that is associated with the possibility of low or high output. The intuition for this result is that the risk neutral principal (who maximizes *expected* profit) assumes all the risk from the risk averse agent.

Which contract should the principal choose in Case 1, the one specifying high effort or the one specifying low effort? The principal chooses the contract that gives a higher expected payoff. The principal's expected payoff from the high effort contract is

$$0.7y_h + 0.3y_l - (\bar{u} + c_h)^2. \tag{6.10}$$

The principal's expected payoff from the low effort contract is

$$0.4y_h + 0.6y_l - (\bar{u} + c_l)^2. \tag{6.11}$$

If $0.3(y_h - y_l) > (\bar{u} + c_h)^2 - (\bar{u} + c_l)^2$, then the high effort contract is better for the principal, otherwise, it is the low effort contract.

Case 2: Effort is unobservable/unverifiable. In this case, the contract cannot specify the effort level of the agent. Observe first that a full insurance contract will lead to low effort by the worker, because, if the worker does not receive any different wage depending on output, then he has no incentive to choose the high effort that is more costly. Note also that the optimal low effort contract from Case 1 above remains feasible.

However, more incentives have to be provided in order to induce high effort. The worker chooses high effort if and only if that gives him at least the same expected utility as low effort:

$$0.7\sqrt{s_h} + 0.3\sqrt{s_l} - c_h \geq 0.4\sqrt{s_h} + 0.6\sqrt{s_l} - c_l, \tag{6.12}$$

or, rearranged,

$$0.3(\sqrt{s_h} - \sqrt{s_l}) \geq c_h - c_l. \tag{6.13}$$

This is called the *incentive constraint* for the worker. The principal's problem is

$$\max 0.7(y_h - s_h) + 0.3(y_l - s_l) \tag{6.14}$$

subject to

$$0.7\sqrt{s_h} + 0.3\sqrt{s_l} - c_h \geq \bar{u} \tag{6.15}$$

$$0.3(\sqrt{s_h} - \sqrt{s_l}) \geq c_h - c_l \tag{6.16}$$

Note that this optimization problem differs from the principal's optimization problem in Case 1 above only by the additional incentive constraint. Differentiating leads to the following first order conditions:

$$-0.7 + 0.7\frac{\lambda}{2\sqrt{s_h}} + 0.3\frac{\mu}{2\sqrt{s_h}} = 0 \tag{6.17}$$

$$-0.3 + 0.3\frac{\lambda}{2\sqrt{s_l}} - 0.3\frac{\mu}{2\sqrt{s_l}} = 0 \tag{6.18}$$

Note that both constraints must be binding. Suppose, to the contrary, that the participation constraint is not binding; then the principal could just decrease s_l a bit, which would leave both the participation constraint and the incentive constraint satisfied and gives the principal a higher expected profit. Suppose next that the incentive constraint is not binding so that it could be ignored. But in this case, the principal's problem is exactly the same as in Case 1 above, which means that it would have the same full insurance solution; since that solution violates the worker's incentive constraint, this cannot be true and so the incentive constraint must be binding.

Solving for s_l and s_h yields $s_l = \frac{(\lambda-\mu)^2}{4}$ and $s_h = \frac{(\lambda+\frac{3}{7}\mu)^2}{4}$. Hence, the wage in the high output state must be larger than in the low output state.

To find the actual values of s_l and s_h, we need to solve both constraints, which is a linear equation system in $\sqrt{s_l}$ and $\sqrt{s_h}$. We find that $s_l = \left(\bar{u} + \frac{7}{3}c_l - \frac{4}{3}c_h\right)^2$ and $s_h = (\bar{u} + 2c_h - c_l)^2$ solves both constraints simultaneously. Note that $s_h > (\bar{u} + c_h)^2$, so if output is high, the worker receives a higher wage than under symmetric information (Case 1). On the other hand, $s_l < (\bar{u} + c_l)^2$, which shows that when output is low, the worker receives less than the optimal wage for the low effort contract (even though he, in equilibrium, exerted high effort!)

The principal's expected profit, when choosing a high effort inducing contract, is

$$0.7y_h + 0.3y_l - \text{expected wage} \tag{6.19}$$

Using the results from above, the expected wage is

$$\begin{aligned} &0.7\left(\bar{u} + 2c_h - c_l\right)^2 + 0.3\left(\bar{u} + \frac{7}{3}c_l - \frac{4}{3}c_h\right)^2 > \\ &\left[0.7\left(\bar{u} + 2c_h - c_l\right) + 0.3\left(\bar{u} + \frac{7}{3}c_l - \frac{4}{3}c_h\right)\right]^2 = \left(\bar{u} + c_h\right)^2 \end{aligned} \tag{6.20}$$

Because $(\bar{u} + c_h)^2$ is the wage that the principal would have to pay if effort were observable and verifiable, the principal is worse off than under observable effort. There are now two possibilities. The first one is that it is still optimal for the principal to use a contract that is meant to induce high effort, even though this is now more expensive (in expectation) than under symmetric information. Second, it is possible that under asymmetric information, the expected profit under the high effort contract goes down by so much that now the low effort contract becomes better for the principal.

In either case, there is a welfare loss due to asymmetric information. Since the agent reaches exactly the same (expected) utility as in Case 1, the difference in the principal's profit measures the welfare loss.

6.4.2 Moral hazard and policy

Moral hazard is an important problem in many economic settings, wherever individuals respond to "insurance" by changing their behavior in a way that is unfavorable for the principal. The following are just some examples:

- Contracts for managers (problems: shirking; "empire building"
- certain insurance contracts (problem: reduced self-protection)
- procurement (e.g., buy fighter jets; if state reimburses costs, firm does not have an incentive to reduce costs)
- Social assistance programs (e.g., unemployment insurance; problem: individuals may reduce effort when looking for a new job)
- taxation (higher marginal tax rate reduces incentives to work longer or increase productivity by education)

Other than in the first two applications of asymmetric information (i.e., adverse selection and signaling), there is no obvious policy recommendation in the case of moral hazard. The state does not have any particular advantage over private parties, when moral hazard is an issue: The state as principal needs to give incentives to its agents, too, pretty much in the same way as private parties.

A Primer in Game Theory

Robert Gibbons

FT Prentice Hall
FINANCIAL TIMES

An imprint of Pearson Education
Harlow, England • London • New York • Boston • San Francisco • Toronto
Sydney • Tokyo • Singapore • Hong Kong • Seoul • Taipei • New Delhi
Cape Town • Madrid • Mexico City • Amsterdam • Munich • Paris • Milan

Chapter 1

Static Games of Complete Information

In this chapter we consider games of the following simple form: first the players simultaneously choose actions; then the players receive payoffs that depend on the combination of actions just chosen. Within the class of such static (or simultaneous-move) games, we restrict attention to games of *complete information*. That is, each player's payoff function (the function that determines the player's payoff from the combination of actions chosen by the players) is common knowledge among all the players. We consider dynamic (or sequential-move) games in Chapters 2 and 4, and games of incomplete information (games in which some player is uncertain about another player's payoff function—as in an auction where each bidder's willingness to pay for the good being sold is unknown to the other bidders) in Chapters 3 and 4.

In Section 1.1 we take a first pass at the two basic issues in game theory: how to describe a game and how to solve the resulting game-theoretic problem. We develop the tools we will use in analyzing static games of complete information, and also the foundations of the theory we will use to analyze richer games in later chapters. We define the *normal-form representation* of a game and the notion of a *strictly dominated strategy*. We show that some games can be solved by applying the idea that rational players do not play strictly dominated strategies, but also that in other games this approach produces a very imprecise prediction about the play of the game (sometimes as imprecise as "anything could

happen"). We then motivate and define *Nash equilibrium*—a solution concept that produces much tighter predictions in a very broad class of games.

In Section 1.2 we analyze four applications, using the tools developed in the previous section: Cournot's (1838) model of imperfect competition, Bertrand's (1883) model of imperfect competition, Farber's (1980) model of final-offer arbitration, and the problem of the commons (discussed by Hume [1739] and others). In each application we first translate an informal statement of the problem into a normal-form representation of the game and then solve for the game's Nash equilibrium. (Each of these applications has a unique Nash equilibrium, but we discuss examples in which this is not true.)

In Section 1.3 we return to theory. We first define the notion of a *mixed strategy*, which we will interpret in terms of one player's uncertainty about what another player will do. We then state and discuss Nash's (1950) Theorem, which guarantees that a Nash equilibrium (possibly involving mixed strategies) exists in a broad class of games. Since we present first basic theory in Section 1.1, then applications in Section 1.2, and finally more theory in Section 1.3, it should be apparent that mastering the additional theory in Section 1.3 is not a prerequisite for understanding the applications in Section 1.2. On the other hand, the ideas of a mixed strategy and the existence of equilibrium do appear (occasionally) in later chapters.

This and each subsequent chapter concludes with problems, suggestions for further reading, and references.

1.1 Basic Theory: Normal-Form Games and Nash Equilibrium

1.1.A Normal-Form Representation of Games

In the normal-form representation of a game, each player simultaneously chooses a strategy, and the combination of strategies chosen by the players determines a payoff for each player. We illustrate the normal-form representation with a classic example — *The Prisoners' Dilemma*. Two suspects are arrested and charged with a crime. The police lack sufficient evidence to convict the suspects, unless at least one confesses. The police hold the suspects in

separate cells and explain the consequences that will follow from the actions they could take. If neither confesses then both will be convicted of a minor offense and sentenced to one month in jail. If both confess then both will be sentenced to jail for six months. Finally, if one confesses but the other does not, then the confessor will be released immediately but the other will be sentenced to nine months in jail—six for the crime and a further three for obstructing justice.

The prisoners' problem can be represented in the accompanying bi-matrix. (Like a matrix, a bi-matrix can have an arbitrary number or rows and columns; "bi" refers to the fact that, in a two-player game, there are two numbers in each cell—the payoffs to the two players.)

		Prisoner 2: Mum	Prisoner 2: Fink
Prisoner 1	Mum	$-1, -1$	$-9, \ 0$
	Fink	$0, -9$	$-6, -6$

The Prisoners' Dilemma

In this game, each player has two strategies available: confess (or fink) and not confess (or be mum). The payoffs to the two players when a particular pair of strategies is chosen are given in the appropriate cell of the bi-matrix. By convention, the payoff to the so-called row player (here, Prisoner 1) is the first payoff given, followed by the payoff to the column player (here, Prisoner 2). Thus, if Prisoner 1 chooses Mum and Prisoner 2 chooses Fink, for example, then Prisoner 1 receives the payoff -9 (representing nine months in jail) and Prisoner 2 receives the payoff 0 (representing immediate release).

We now turn to the general case. The *normal-form representation* of a game specifies: (1) the players in the game, (2) the strategies available to each player, and (3) the payoff received by each player for each combination of strategies that could be chosen by the players. We will often discuss an n-player game in which the players are numbered from 1 to n and an arbitrary player is called player i. Let S_i denote the set of strategies available to player i (called i's *strategy space*), and let s_i denote an arbitrary member of this set. (We will occasionally write $s_i \in S_i$ to indicate that the

strategy s_i is a member of the set of strategies S_i.) Let $(s_1, \ldots, s_n)$ denote a combination of strategies, one for each player, and let u_i denote player i's *payoff function*: $u_i(s_1, \ldots, s_n)$ is the payoff to player i if the players choose the strategies $(s_1, \ldots, s_n)$. Collecting all of this information together, we have:

Definition *The* ***normal-form representation*** *of an n-player game specifies the players' strategy spaces $S_1, \ldots, S_n$ and their payoff functions $u_1, \ldots, u_n$. We denote this game by $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$.*

Although we stated that in a normal-form game the players choose their strategies simultaneously, this does not imply that the parties necessarily *act* simultaneously: it suffices that each choose his or her action without knowledge of the others' choices, as would be the case here if the prisoners reached decisions at arbitrary times while in their separate cells. Furthermore, although in this chapter we use normal-form games to represent only static games in which the players all move without knowing the other players' choices, we will see in Chapter 2 that normal-form representations can be given for sequential-move games, but also that an alternative—the *extensive-form* representation of the game—is often a more convenient framework for analyzing dynamic issues.

1.1.B Iterated Elimination of Strictly Dominated Strategies

Having described one way to represent a game, we now take a first pass at describing how to solve a game-theoretic problem. We start with the Prisoners' Dilemma because it is easy to solve, using only the idea that a rational player will not play a strictly dominated strategy.

In the Prisoners' Dilemma, if one suspect is going to play Fink, then the other would prefer to play Fink and so be in jail for six months rather than play Mum and so be in jail for nine months. Similarly, if one suspect is going to play Mum, then the other would prefer to play Fink and so be released immediately rather than play Mum and so be in jail for one month. Thus, for prisoner i, playing Mum is dominated by playing Fink—for each strategy that prisoner j could choose, the payoff to prisoner i from playing Mum is less than the payoff to i from playing Fink. (The same would be true in any bi-matrix in which the payoffs 0, −1, −6, and −9 above were replaced with payoffs T, R, P, and S, respectively, provided that $T > R > P > S$ so as to capture the ideas of temptation, reward, punishment, and sucker payoffs.) More generally:

Definition *In the normal-form game $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$, let s_i' and s_i'' be feasible strategies for player i (i.e., s_i' and s_i'' are members of S_i). Strategy s_i' is* ***strictly dominated*** *by strategy s_i'' if for each feasible combination of the other players' strategies, i's payoff from playing s_i' is strictly less than i's payoff from playing s_i'':*

$$u_i(s_1, \ldots, s_{i-1}, s_i', s_{i+1}, \ldots, s_n) < u_i(s_1, \ldots, s_{i-1}, s_i'', s_{i+1}, \ldots, s_n) \quad \text{(DS)}$$

for each $(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ that can be constructed from the other players' strategy spaces $S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n$.

Rational players do not play strictly dominated strategies, because there is no belief that a player could hold (about the strategies the other players will choose) such that it would be optimal to play such a strategy.[1] Thus, in the Prisoners' Dilemma, a rational player will choose Fink, so (Fink, Fink) will be the outcome reached by two rational players, even though (Fink, Fink) results in worse payoffs for both players than would (Mum, Mum). Because the Prisoners' Dilemma has many applications (including the arms race and the free-rider problem in the provision of public goods), we will return to variants of the game in Chapters 2 and 4. For now, we focus instead on whether the idea that rational players do not play strictly dominated strategies can lead to the solution of other games.

Consider the abstract game in Figure 1.1.1.[2] Player 1 has two strategies and player 2 has three: $S_1 = \{\text{Up}, \text{Down}\}$ and $S_2 = \{\text{Left}, \text{Middle}, \text{Right}\}$. For player 1, neither Up nor Down is strictly

[1] A complementary question is also of interest: if there is no belief that player i could hold (about the strategies the other players will choose) such that it would be optimal to play the strategy s_i, can we conclude that there must be another strategy that strictly dominates s_i? The answer is "yes," *provided* that we adopt appropriate definitions of "belief" and "another strategy," both of which involve the idea of mixed strategies to be introduced in Section 1.3.A.

[2] Most of this book considers economic applications rather than abstract examples, both because the applications are of interest in their own right and because, for many readers, the applications are often a useful way to explain the underlying theory. When introducing some of the basic theoretical ideas, however, we will sometimes resort to abstract examples that have no natural economic interpretation.

		Player 2		
		Left	Middle	Right
Player 1	Up	1,0	1,2	0,1
	Down	0,3	0,1	2,0

Figure 1.1.1.

dominated: Up is better than Down if 2 plays Left (because $1 > 0$), but Down is better than Up if 2 plays Right (because $2 > 0$). For player 2, however, Right is strictly dominated by Middle (because $2 > 1$ and $1 > 0$), so a rational player 2 will not play Right. Thus, if player 1 knows that player 2 is rational then player 1 can eliminate Right from player 2's strategy space. That is, if player 1 knows that player 2 is rational then player 1 can play the game in Figure 1.1.1 *as if* it were the game in Figure 1.1.2.

		Player 2	
		Left	Middle
Player 1	Up	1,0	1,2
	Down	0,3	0,1

Figure 1.1.2.

In Figure 1.1.2, Down is now strictly dominated by Up for player 1, so if player 1 is rational (and player 1 knows that player 2 is rational, so that the game in Figure 1.1.2 applies) then player 1 will not play Down. Thus, if player 2 knows that player 1 is rational, *and* player 2 knows that player 1 knows that player 2 is rational (so that player 2 knows that Figure 1.1.2 applies), then player 2 can eliminate Down from player 1's strategy space, leaving the game in Figure 1.1.3. But now Left is strictly dominated by Middle for player 2, leaving (Up, Middle) as the outcome of the game.

This process is called *iterated elimination of strictly dominated strategies*. Although it is based on the appealing idea that rational players do not play strictly dominated strategies, the process has two drawbacks. First, each step requires a further assumption

		Player 2	
		Left	Middle
Player 1	Up	1,0	1,2

Figure 1.1.3.

about what the players know about each other's rationality. If we want to be able to apply the process for an arbitrary number of steps, we need to assume that it is *common knowledge* that the players are rational. That is, we need to assume not only that all the players are rational, but also that all the players know that all the players are rational, and that all the players know that all the players know that all the players are rational, and so on, *ad infinitum*. (See Aumann [1976] for the formal definition of common knowledge.)

The second drawback of iterated elimination of strictly dominated strategies is that the process often produces a very imprecise prediction about the play of the game. Consider the game in Figure 1.1.4, for example. In this game there are no strictly dominated strategies to be eliminated. (Since we have not motivated this game in the slightest, it may appear arbitrary, or even pathological. See the case of three or more firms in the Cournot model in Section 1.2.A for an economic application in the same spirit.) Since all the strategies in the game survive iterated elimination of strictly dominated strategies, the process produces no prediction whatsoever about the play of the game.

	L	*C*	*R*
T	0,4	4,0	5,3
M	4,0	0,4	5,3
B	3,5	3,5	6,6

Figure 1.1.4.

We turn next to Nash equilibrium—a solution concept that produces much tighter predictions in a very broad class of games. We show that Nash equilibrium is a stronger solution concept

than iterated elimination of strictly dominated strategies, in the sense that the players' strategies in a Nash equilibrium always survive iterated elimination of strictly dominated strategies, but the converse is not true. In subsequent chapters we will argue that in richer games even Nash equilibrium produces too imprecise a prediction about the play of the game, so we will define still-stronger notions of equilibrium that are better suited for these richer games.

1.1.C Motivation and Definition of Nash Equilibrium

One way to motivate the definition of Nash equilibrium is to argue that if game theory is to provide a unique solution to a game-theoretic problem then the solution must be a Nash equilibrium, in the following sense. Suppose that game theory makes a unique prediction about the strategy each player will choose. In order for this prediction to be correct, it is necessary that each player be willing to choose the strategy predicted by the theory. Thus, each player's predicted strategy must be that player's best response to the predicted strategies of the other players. Such a prediction could be called *strategically stable* or *self-enforcing*, because no single player wants to deviate from his or her predicted strategy. We will call such a prediction a Nash equilibrium:

Definition *In the n-player normal-form game* $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$, *the strategies* $(s_1^*, \ldots, s_n^*)$ *are a* ***Nash equilibrium*** *if, for each player* i, s_i^* *is (at least tied for) player* i*'s best response to the strategies specified for the* $n - 1$ *other players,* $(s_1^*, \ldots, s_{i-1}^*, s_{i+1}^*, \ldots, s_n^*)$*:*

$$u_i(s_1^*, \ldots, s_{i-1}^*, s_i^*, s_{i+1}^*, \ldots, s_n^*) \geq u_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_n^*) \qquad (NE)$$

for every feasible strategy s_i *in* S_i*; that is,* s_i^* *solves*

$$\max_{s_i \in S_i} u_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_n^*).$$

To relate this definition to its motivation, suppose game theory offers the strategies $(s_1', \ldots, s_n')$ as the solution to the normal-form game $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$. Saying that $(s_1', \ldots, s_n')$ is *not* a Nash equilibrium of G is equivalent to saying that there exists some player i such that s_i' is *not* a best response to $(s_1', \ldots, s_{i-1}', s_{i+1}', \ldots, s_n')$. That is, there exists some s_i'' in S_i such that

$$u_i(s_1', \ldots, s_{i-1}', s_i', s_{i+1}', \ldots, s_n') < u_i(s_1', \ldots, s_{i-1}', s_i'', s_{i+1}', \ldots, s_n').$$

Thus, if the theory offers the strategies $(s_1', \ldots, s_n')$ as the solution but these strategies are not a Nash equilibrium, then at least one player will have an incentive to deviate from the theory's prediction, so the theory will be falsified by the actual play of the game. A closely related motivation for Nash equilibrium involves the idea of convention: if a convention is to develop about how to play a given game then the strategies prescribed by the convention must be a Nash equilibrium, else at least one player will not abide by the convention.

To be more concrete, we now solve a few examples. Consider the three normal-form games already described—the Prisoners' Dilemma and Figures 1.1.1 and 1.1.4. A brute-force approach to finding a game's Nash equilibria is simply to check whether each possible combination of strategies satisfies condition (NE) in the definition.[3] In a two-player game, this approach begins as follows: for each player, and for each feasible strategy for that player, determine the other player's best response to that strategy. Figure 1.1.5 does this for the game in Figure 1.1.4 by underlining the payoff to player j's best response to each of player i's feasible strategies. If the column player were to play L, for instance, then the row player's best response would be M, since 4 exceeds 3 and 0, so the row player's payoff of 4 in the (M, L) cell of the bi-matrix is underlined.

A pair of strategies satisfies condition (NE) if each player's strategy is a best response to the other's—that is, if both payoffs are underlined in the corresponding cell of the bi-matrix. Thus, (B, R) is the only strategy pair that satisfies (NE); likewise for (Fink, Fink) in the Prisoners' Dilemma and (Up, Middle) in

[3] In Section 1.3.A we will distinguish between pure and mixed strategies. We will then see that the definition given here describes *pure-strategy* Nash equilibria, but that there can also be *mixed-strategy* Nash equilibria. Unless explicitly noted otherwise, all references to Nash equilibria in this section are to pure-strategy Nash equilibria.

	L	C	R
T	0,<u>4</u>	<u>4</u>,0	5,3
M	<u>4</u>,0	0,<u>4</u>	5,3
B	3,5	3,5	<u>6</u>,<u>6</u>

Figure 1.1.5.

Figure 1.1.1. These strategy pairs are the unique Nash equilibria of these games.[4]

We next address the relation between Nash equilibrium and iterated elimination of strictly dominated strategies. Recall that the Nash equilibrium strategies in the Prisoners' Dilemma and Figure 1.1.1—(Fink, Fink) and (Up, Middle), respectively—are the only strategies that survive iterated elimination of strictly dominated strategies. This result can be generalized: if iterated elimination of strictly dominated strategies eliminates all but the strategies $(s_1^*,\ldots,s_n^*)$, then these strategies are the unique Nash equilibrium of the game. (See Appendix 1.1.C for a proof of this claim.) Since iterated elimination of strictly dominated strategies frequently does *not* eliminate all but a single combination of strategies, however, it is of more interest that Nash equilibrium is a stronger solution concept than iterated elimination of strictly dominated strategies, in the following sense. If the strategies $(s_1^*,\ldots,s_n^*)$ are a Nash equilibrium then they survive iterated elimination of strictly dominated strategies (again, see the Appendix for a proof), but there can be strategies that survive iterated elimination of strictly dominated strategies but are not part of any Nash equilibrium. To see the latter, recall that in Figure 1.1.4 Nash equilibrium gives the unique prediction (B, R), whereas iterated elimination of strictly dominated strategies gives the maximally imprecise prediction: no strategies are eliminated; anything could happen.

Having shown that Nash equilibrium is a stronger solution concept than iterated elimination of strictly dominated strategies, we must now ask whether Nash equilibrium is too strong a solution concept. That is, can we be sure that a Nash equilibrium

[4]This statement is correct even if we do not restrict attention to pure-strategy Nash equilibrium, because no mixed-strategy Nash equilibria exist in these three games. See Problem 1.10.

exists? Nash (1950) showed that in any finite game (i.e., a game in which the number of players n and the strategy sets $S_1,\ldots,S_n$ are all finite) there exists at least one Nash equilibrium. (This equilibrium may involve mixed strategies, which we will discuss in Section 1.3.A; see Section 1.3.B for a precise statement of Nash's Theorem.) Cournot (1838) proposed the same notion of equilibrium in the context of a particular model of duopoly and demonstrated (by construction) that an equilibrium exists in that model; see Section 1.2.A. In every application analyzed in this book, we will follow Cournot's lead: we will demonstrate that a Nash (or stronger) equilibrium exists by constructing one. In some of the theoretical sections, however, we will rely on Nash's Theorem (or its analog for stronger equilibrium concepts) and simply assert that an equilibrium exists.

We conclude this section with another classic example—*The Battle of the Sexes*. This example shows that a game can have multiple Nash equilibria, and also will be useful in the discussions of mixed strategies in Sections 1.3.B and 3.2.A. In the traditional exposition of the game (which, it will be clear, dates from the 1950s), a man and a woman are trying to decide on an evening's entertainment; we analyze a gender-neutral version of the game. While at separate workplaces, Pat and Chris must choose to attend either the opera or a prize fight. Both players would rather spend the evening together than apart, but Pat would rather they be together at the prize fight while Chris would rather they be together at the opera, as represented in the accompanying bi-matrix.

		Pat: Opera	Pat: Fight
Chris	Opera	2, 1	0, 0
	Fight	0, 0	1, 2

The Battle of the Sexes

Both (Opera, Opera) and (Fight, Fight) are Nash equilibria.

We argued above that if game theory is to provide a unique solution to a game then the solution must be a Nash equilibrium. This argument ignores the possibility of games in which game theory does not provide a unique solution. We also argued that

if a convention is to develop about how to play a given game, then the strategies prescribed by the convention must be a Nash equilibrium, but this argument similarly ignores the possibility of games for which a convention will not develop. In some games with multiple Nash equilibria one equilibrium stands out as the compelling solution to the game. (Much of the theory in later chapters is an effort to identify such a compelling equilibrium in different classes of games.) Thus, the existence of multiple Nash equilibria is not a problem in and of itself. In the Battle of the Sexes, however, (Opera, Opera) and (Fight, Fight) seem equally compelling, which suggests that there may be games for which game theory does not provide a unique solution and no convention will develop.[5] In such games, Nash equilibrium loses much of its appeal as a prediction of play.

Appendix 1.1.C

This appendix contains proofs of the following two Propositions, which were stated informally in Section 1.1.C. Skipping these proofs will not substantially hamper one's understanding of later material. For readers not accustomed to manipulating formal definitions and constructing proofs, however, mastering these proofs will be a valuable exercise.

Proposition A *In the n-player normal-form game* $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$, *if iterated elimination of strictly dominated strategies eliminates all but the strategies* $(s_1^*, \ldots, s_n^*)$, *then these strategies are the unique Nash equilibrium of the game.*

Proposition B *In the n-player normal-form game* $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$, *if the strategies* $(s_1^*, \ldots, s_n^*)$ *are a Nash equilibrium, then they survive iterated elimination of strictly dominated strategies.*

[5] In Section 1.3.B we describe a third Nash equilibrium of the Battle of the Sexes (involving mixed strategies). Unlike (Opera, Opera) and (Fight, Fight), this third equilibrium has symmetric payoffs, as one might expect from the unique solution to a symmetric game; on the other hand, the third equilibrium is also inefficient, which may work against its development as a convention. Whatever one's judgment about the Nash equilibria in the Battle of the Sexes, however, the broader point remains: there may be games in which game theory does not provide a unique solution and no convention will develop.

Since Proposition B is simpler to prove, we begin with it, to warm up. The argument is by contradiction. That is, we will assume that one of the strategies in a Nash equilibrium is eliminated by iterated elimination of strictly dominated strategies, and then we will show that a contradiction would result if this assumption were true, thereby proving that the assumption must be false.

Suppose that the strategies $(s_1^*, \ldots, s_n^*)$ are a Nash equilibrium of the normal-form game $G = \{S_1, \ldots, S_n; u_1, \ldots, u_n\}$, but suppose also that (perhaps after some strategies other than $(s_1^*, \ldots, s_n^*)$ have been eliminated) s_i^* is the first of the strategies $(s_1^*, \ldots, s_n^*)$ to be eliminated for being strictly dominated. Then there must exist a strategy s_i'' that has not yet been eliminated from S_i that strictly dominates s_i^*. Adapting (DS), we have

$$u_i(s_1, \ldots, s_{i-1}, s_i^*, s_{i+1}, \ldots, s_n) < u_i(s_1, \ldots, s_{i-1}, s_i'', s_{i+1}, \ldots, s_n) \qquad (1.1.1)$$

for each $(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ that can be constructed from the strategies that have not yet been eliminated from the other players' strategy spaces. Since s_i^* is the first of the equilibrium strategies to be eliminated, the other players' equilibrium strategies have not yet been eliminated, so one of the implications of (1.1.1) is

$$u_i(s_1^*, \ldots, s_{i-1}^*, s_i^*, s_{i+1}^*, \ldots, s_n^*) < u_i(s_1^*, \ldots, s_{i-1}^*, s_i'', s_{i+1}^*, \ldots, s_n^*). \qquad (1.1.2)$$

But (1.1.2) is contradicted by (NE): s_i^* must be a best response to $(s_1^*, \ldots, s_{i-1}^*, s_{i+1}^*, \ldots, s_n^*)$, so there cannot exist a strategy s_i'' that strictly dominates s_i^*. This contradiction completes the proof.

Having proved Proposition B, we have already proved part of Proposition A: all we need to show is that if iterated elimination of dominated strategies eliminates all but the strategies $(s_1^*, \ldots, s_n^*)$ then these strategies are a Nash equilibrium; by Proposition B, any other Nash equilibria would also have survived, so this equilibrium must be unique. We assume that G is finite.

The argument is again by contradiction. Suppose that iterated elimination of dominated strategies eliminates all but the strategies $(s_1^*, \ldots, s_n^*)$ but these strategies are not a Nash equilibrium. Then there must exist some player i and some feasible strategy s_i in S_i such that (NE) fails, but s_i must have been strictly dominated by some other strategy s_i' at some stage of the process. The formal

statements of these two observations are: there exists s_i in S_i such that

$$u_i(s_1^*, \ldots, s_{i-1}^*, s_i^*, s_{i+1}^*, \ldots, s_n^*) < u_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_n^*); \quad (1.1.3)$$

and there exists s_i' in the set of player i's strategies remaining at some stage of the process such that

$$u_i(s_1, \ldots, s_{i-1}, s_i, s_{i+1}, \ldots, s_n) < u_i(s_1, \ldots, s_{i-1}, s_i', s_{i+1}, \ldots, s_n) \quad (1.1.4)$$

for each $(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ that can be constructed from the strategies remaining in the other players' strategy spaces at that stage of the process. Since the other players' strategies $(s_1^*, \ldots, s_{i-1}^*, s_{i+1}^*, \ldots, s_n^*)$ are never eliminated, one of the implications of (1.1.4) is

$$u_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_n^*) < u_i(s_1^*, \ldots, s_{i-1}^*, s_i', s_{i+1}^*, \ldots, s_n^*). \quad (1.1.5)$$

If $s_i' = s_i^*$ (that is, if s_i^* is the strategy that strictly dominates s_i) then (1.1.5) contradicts (1.1.3), in which case the proof is complete. If $s_i' \neq s_i^*$ then some other strategy s_i'' must later strictly dominate s_i', since s_i' does not survive the process. Thus, inequalities analogous to (1.1.4) and (1.1.5) hold with s_i' and s_i'' replacing s_i and s_i', respectively. Once again, if $s_i'' = s_i^*$ then the proof is complete; otherwise, two more analogous inequalities can be constructed. Since s_i^* is the only strategy from S_i to survive the process, repeating this argument (in a finite game) eventually completes the proof.

1.2 Applications

1.2.A Cournot Model of Duopoly

As noted in the previous section, Cournot (1838) anticipated Nash's definition of equilibrium by over a century (but only in the context of a particular model of duopoly). Not surprisingly, Cournot's work is one of the classics of game theory; it is also one of the cornerstones of the theory of industrial organization. We consider a very simple version of Cournot's model here, and return to variations on the model in each subsequent chapter. In this section we use the model to illustrate: (a) the translation of an informal statement of a problem into a normal-form representation of a game; (b) the computations involved in solving for the game's Nash equilibrium; and (c) iterated elimination of strictly dominated strategies.

Let q_1 and q_2 denote the quantities (of a homogeneous product) produced by firms 1 and 2, respectively. Let $P(Q) = a - Q$ be the market-clearing price when the aggregate quantity on the market is $Q = q_1 + q_2$. (More precisely, $P(Q) = a - Q$ for $Q < a$, and $P(Q) = 0$ for $Q \geq a$.) Assume that the total cost to firm i of producing quantity q_i is $C_i(q_i) = cq_i$. That is, there are no fixed costs and the marginal cost is constant at c, where we assume $c < a$. Following Cournot, suppose that the firms choose their quantities simultaneously.[6]

In order to find the Nash equilibrium of the Cournot game, we first translate the problem into a normal-form game. Recall from the previous section that the normal-form representation of a game specifies: (1) the players in the game, (2) the strategies available to each player, and (3) the payoff received by each player for each combination of strategies that could be chosen by the players. There are of course two players in any duopoly game—the two firms. In the Cournot model, the strategies available to each firm are the different quantities it might produce. We will assume that output is continuously divisible. Naturally, negative outputs are not feasible. Thus, each firm's strategy space can be represented as $S_i = [0, \infty)$, the nonnegative real numbers, in which case a typical strategy s_i is a quantity choice, $q_i \geq 0$. One could argue that extremely large quantities are not feasible and so should not be included in a firm's strategy space. Because $P(Q) = 0$ for $Q \geq a$, however, neither firm will produce a quantity $q_i > a$.

It remains to specify the payoff to firm i as a function of the strategies chosen by it and by the other firm, and to define and

[6] We discuss Bertrand's (1883) model, in which firms choose prices rather than quantities, in Section 1.2.B, and Stackelberg's (1934) model, in which firms choose quantities but one firm chooses before (and is observed by) the other, in Section 2.1.B. Finally, we discuss Friedman's (1971) model, in which the interaction described in Cournot's model occurs repeatedly over time, in Section 2.3.C.

solve for equilibrium. We assume that the firm's payoff is simply its profit. Thus, the payoff $u_i(s_i, s_j)$ in a general two-player game in normal form can be written here as[7]

$$\pi_i(q_i, q_j) = q_i[P(q_i + q_j) - c] = q_i[a - (q_i + q_j) - c].$$

Recall from the previous section that in a two-player game in normal form, the strategy pair (s_1^*, s_2^*) is a Nash equilibrium if, for each player i,

$$u_i(s_i^*, s_j^*) \geq u_i(s_i, s_j^*) \qquad \text{(NE)}$$

for every feasible strategy s_i in S_i. Equivalently, for each player i, s_i^* must solve the optimization problem

$$\max_{s_i \in S_i} u_i(s_i, s_j^*).$$

In the Cournot duopoly model, the analogous statement is that the quantity pair (q_1^*, q_2^*) is a Nash equilibrium if, for each firm i, q_i^* solves

$$\max_{0 \leq q_i < \infty} \pi_i(q_i, q_j^*) = \max_{0 \leq q_i < \infty} q_i[a - (q_i + q_j^*) - c].$$

Assuming $q_j^* < a - c$ (as will be shown to be true), the first-order condition for firm i's optimization problem is both necessary and sufficient; it yields

$$q_i = \frac{1}{2}(a - q_j^* - c). \qquad (1.2.1)$$

Thus, if the quantity pair (q_1^*, q_2^*) is to be a Nash equilibrium, the firms' quantity choices must satisfy

$$q_1^* = \frac{1}{2}(a - q_2^* - c)$$

and

$$q_2^* = \frac{1}{2}(a - q_1^* - c).$$

[7]Note that we have changed the notation slightly by writing $u_i(s_i, s_j)$ rather than $u_i(s_1, s_2)$. Both expressions represent the payoff to player i as a function of the strategies chosen by all the players. We will use these expressions (and their n-player analogs) interchangeably.

Solving this pair of equations yields

$$q_1^* = q_2^* = \frac{a - c}{3},$$

which is indeed less than $a - c$, as assumed.

The intuition behind this equilibrium is simple. Each firm would of course like to be a monopolist in this market, in which case it would choose q_i to maximize $\pi_i(q_i, 0)$—it would produce the monopoly quantity $q_m = (a - c)/2$ and earn the monopoly profit $\pi_i(q_m, 0) = (a - c)^2/4$. Given that there are two firms, aggregate profits for the duopoly would be maximized by setting the aggregate quantity $q_1 + q_2$ equal to the monopoly quantity q_m, as would occur if $q_i = q_m/2$ for each i, for example. The problem with this arrangement is that each firm has an incentive to deviate: because the monopoly quantity is low, the associated price $P(q_m)$ is high, and at this price each firm would like to increase its quantity, in spite of the fact that such an increase in production drives down the market-clearing price. (To see this formally, use (1.2.1) to check that $q_m/2$ is *not* firm 2's best response to the choice of $q_m/2$ by firm 1.) In the Cournot equilibrium, in contrast, the aggregate quantity is higher, so the associated price is lower, so the temptation to increase output is reduced—reduced by just enough that each firm is just deterred from increasing its output by the realization that the market-clearing price will fall. See Problem 1.4 for an analysis of how the presence of n oligopolists affects this equilibrium trade-off between the temptation to increase output and the reluctance to reduce the market-clearing price.

Rather than solving for the Nash equilibrium in the Cournot game algebraically, one could instead proceed graphically, as follows. Equation (1.2.1) gives firm i's best response to firm j's *equilibrium* strategy, q_j^*. Analogous reasoning leads to firm 2's best response to an *arbitrary* strategy by firm 1 and firm 1's best response to an arbitrary strategy by firm 2. Assuming that firm 1's strategy satisfies $q_1 < a - c$, firm 2's best response is

$$R_2(q_1) = \frac{1}{2}(a - q_1 - c);$$

likewise, if $q_2 < a - c$ then firm 1's best response is

$$R_1(q_2) = \frac{1}{2}(a - q_2 - c).$$

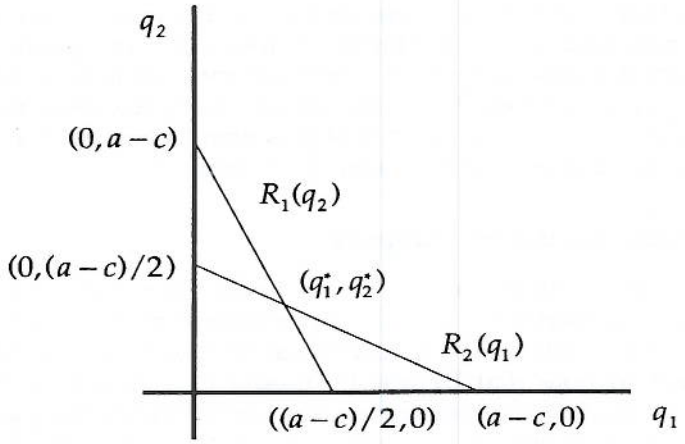


Figure 1.2.1.

As shown in Figure 1.2.1, these two best-response functions intersect only once, at the equilibrium quantity pair (q_1^*, q_2^*).

A third way to solve for this Nash equilibrium is to apply the process of iterated elimination of strictly dominated strategies. This process yields a unique solution—which, by Proposition A in Appendix 1.1.C, must be the Nash equilibrium (q_1^*, q_2^*). The complete process requires an infinite number of steps, each of which eliminates a fraction of the quantities remaining in each firm's strategy space; we discuss only the first two steps. First, the monopoly quantity $q_m = (a-c)/2$ strictly dominates any higher quantity. That is, for any $x > 0$, $\pi_i(q_m, q_j) > \pi_i(q_m + x, q_j)$ for all $q_j \geq 0$. To see this, note that if $Q = q_m + x + q_j < a$, then

$$\pi_i(q_m, q_j) = \frac{a-c}{2}\left[\frac{a-c}{2} - q_j\right]$$

and

$$\pi_i(q_m + x, q_j) = \left[\frac{a-c}{2} + x\right]\left[\frac{a-c}{2} - x - q_j\right] = \pi_i(q_m, q_j) - x(x + q_j),$$

and if $Q = q_m + x + q_j \geq a$, then $P(Q) = 0$, so producing a smaller

quantity raises profit. Second, given that quantities exceeding q_m have been eliminated, the quantity $(a-c)/4$ strictly dominates any lower quantity. That is, for any x between zero and $(a-c)/4$, $\pi_i[(a-c)/4, q_j] > \pi_i[(a-c)/4 - x, q_j]$ for all q_j between zero and $(a-c)/2$. To see this, note that

$$\pi_i\left(\frac{a-c}{4}, q_j\right) = \frac{a-c}{4}\left[\frac{3(a-c)}{4} - q_j\right]$$

and

$$\begin{aligned}\pi_i\left(\frac{a-c}{4} - x, q_j\right) &= \left[\frac{a-c}{4} - x\right]\left[\frac{3(a-c)}{4} + x - q_j\right] \\ &= \pi_i(q_m, q_j) - x\left[\frac{a-c}{2} + x - q_j\right].\end{aligned}$$

After these two steps, the quantities remaining in each firm's strategy space are those in the interval between $(a-c)/4$ and $(a-c)/2$. Repeating these arguments leads to ever-smaller intervals of remaining quantities. In the limit, these intervals converge to the single point $q_i^* = (a-c)/3$.

Iterated elimination of strictly dominated strategies can also be described graphically, by using the observation (from footnote 1; see also the discussion in Section 1.3.A) that a strategy is strictly dominated if and only if there is no belief about the other players' choices for which the strategy is a best response. Since there are only two firms in this model, we can restate this observation as: a quantity q_i is strictly dominated if and only if there is no belief about q_j such that q_i is firm i's best response. We again discuss only the first two steps of the iterative process. First, it is never a best response for firm i to produce more than the monopoly quantity, $q_m = (a-c)/2$. To see this, consider firm 2's best-response function, for example: in Figure 1.2.1, $R_2(q_1)$ equals q_m when $q_1 = 0$ and declines as q_1 increases. Thus, for any $q_j \geq 0$, if firm i believes that firm j will choose q_j, then firm i's best response is less than or equal to q_m; there is no q_j such that firm i's best response exceeds q_m. Second, given this upper bound on firm j's quantity, we can derive a lower bound on firm i's best response: if $q_j \leq (a-c)/2$, then $R_i(q_j) \geq (a-c)/4$, as shown for firm 2's best response in Figure 1.2.2.[8]

[8]These two arguments are slightly incomplete because we have not analyzed

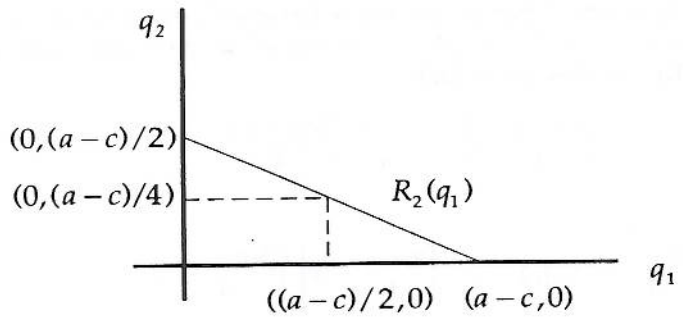


Figure 1.2.2.

As before, repeating these arguments leads to the single quantity $q_i^* = (a - c)/3$.

We conclude this section by changing the Cournot model so that iterated elimination of strictly dominated strategies does *not* yield a unique solution. To do this, we simply add one or more firms to the existing duopoly. We will see that the first of the two steps discussed in the duopoly case continues to hold, but that the process ends there. Thus, when there are more than two firms, iterated elimination of strictly dominated strategies yields only the imprecise prediction that each firm's quantity will not exceed the monopoly quantity (much as in Figure 1.1.4, where no strategies were eliminated by this process).

For concreteness, we consider the three-firm case. Let Q_{-i} denote the sum of the quantities chosen by the firms other than i, and let $\pi_i(q_i, Q_{-i}) = q_i(a - q_i - Q_{-i} - c)$ provided $q_i + Q_{-i} < a$ (whereas $\pi_i(q_i, Q_{-i}) = -cq_i$ if $q_i + Q_{-i} \geq a$). It is again true that the monopoly quantity $q_m = (a - c)/2$ strictly dominates any higher quantity. That is, for any $x > 0$, $\pi_i(q_m, Q_{-i}) > \pi_i(q_m + x, Q_{-i})$ for all $Q_{-i} \geq 0$, just as in the first step in the duopoly case. Since

firm i's best response when firm i is uncertain about q_j. Suppose firm i is uncertain about q_j but believes that the expected value of q_j is $E(q_j)$. Because $\pi_i(q_i, q_j)$ is linear in q_j, firm i's best response when it is uncertain in this way simply equals its best response when it is certain that firm j will choose $E(q_j)$—a case covered in the text.

there are two firms other than firm i, however, all we can say about Q_{-i} is that it is between zero and $a - c$, because q_j and q_k are between zero and $(a - c)/2$. But this implies that no quantity $q_i \geq 0$ is strictly dominated for firm i, because for each q_i between zero and $(a - c)/2$ there exists a value of Q_{-i} between zero and $a - c$ (namely, $Q_{-i} = a - c - 2q_i$) such that q_i is firm i's best response to Q_{-i}. Thus, no further strategies can be eliminated.

1.2.B Bertrand Model of Duopoly

We next consider a different model of how two duopolists might interact, based on Bertrand's (1883) suggestion that firms actually choose prices, rather than quantities as in Cournot's model. It is important to note that Bertrand's model is a *different game* than Cournot's model: the strategy spaces are different, the payoff functions are different, and (as will become clear) the behavior in the Nash equilibria of the two models is different. Some authors summarize these differences by referring to the Cournot and Bertrand equilibria. Such usage may be misleading: it refers to the difference between the Cournot and Bertrand games, and to the difference between the equilibrium behavior in these games, *not* to a difference in the equilibrium concept used in the games. *In both games, the equilibrium concept used is the Nash equilibrium defined in the previous section.*

We consider the case of differentiated products. (See Problem 1.7 for the case of homogeneous products.) If firms 1 and 2 choose prices p_1 and p_2, respectively, the quantity that consumers demand from firm i is

$$q_i(p_i, p_j) = a - p_i + bp_j,$$

where $b > 0$ reflects the extent to which firm i's product is a substitute for firm j's product. (This is an unrealistic demand function because demand for firm i's product is positive even when firm i charges an arbitrarily high price, provided firm j also charges a high enough price. As will become clear, the problem makes sense only if $b < 2$.) As in our discussion of the Cournot model, we assume that there are no fixed costs of production and that marginal costs are constant at c, where $c < a$, and that the firms act (i.e., choose their prices) simultaneously.

As before, the first task in the process of finding the Nash equilibrium is to translate the problem into a normal-form game. There

Chapter 2

Dynamic Games of Complete Information

In this chapter we introduce dynamic games. We again restrict attention to games with complete information (i.e., games in which the players' payoff functions are common knowledge); see Chapter 3 for the introduction to games of incomplete information. In Section 2.1 we analyze dynamic games that have not only complete but also *perfect information*, by which we mean that at each move in the game the player with the move knows the full history of the play of the game thus far. In Sections 2.2 through 2.4 we consider games of complete but imperfect information: at some move the player with the move does not know the history of the game.

The central issue in all dynamic games is credibility. As an example of a noncredible threat, consider the following two-move game. First, player 1 chooses between giving player 2 $1,000 and giving player 2 nothing. Second, player 2 observes player 1's move and then chooses whether or not to explode a grenade that will kill both players. Suppose player 2 threatens to explode the grenade unless player 1 pays the $1,000. If player 1 believes the threat, then player 1's best response is to pay the $1,000. But player 1 should not believe the threat, because it is noncredible: if player 2 were given the opportunity to carry out the threat,

player 2 would choose not to carry it out. Thus, player 1 should pay player 2 nothing.[1]

In Section 2.1 we analyze the following class of dynamic games of complete and perfect information: first player 1 moves, then player 2 observes player 1's move, then player 2 moves and the game ends. The grenade game belongs to this class, as do Stackelberg's (1934) model of duopoly and Leontief's (1946) model of wage and employment determination in a unionized firm. We define the *backwards-induction outcome* of such games and briefly discuss its relation to Nash equilibrium (deferring the main discussion of this relation until Section 2.4). We solve for this outcome in the Stackelberg and Leontief models. We also derive the analogous outcome in Rubinstein's (1982) bargaining model, although this game has a potentially infinite sequence of moves and so does not belong to the above class of games.

In Section 2.2 we enrich the class of games analyzed in the previous section: first players 1 and 2 move simultaneously, then players 3 and 4 observe the moves chosen by 1 and 2, then players 3 and 4 move simultaneously and the game ends. As will be explained in Section 2.4, the simultaneity of moves here means that these games have imperfect information. We define the *subgame-perfect outcome* of such games, which is the natural extension of backwards induction to these games. We solve for this outcome in Diamond and Dybvig's (1983) model of bank runs, in a model of tariffs and imperfect international competition, and in Lazear and Rosen's (1981) model of tournaments.

In Section 2.3 we study *repeated games*, in which a fixed group of players plays a given game repeatedly, with the outcomes of all previous plays observed before the next play begins. The theme of the analysis is that (credible) threats and promises about future behavior can influence current behavior. We define *subgame-perfect Nash equilibrium* for repeated games and relate it to the backwards-induction and subgame-perfect outcomes defined in Sections 2.1 and 2.2. We state and prove the Folk Theorem for infinitely repeated games, and we analyze Friedman's (1971) model of collusion between Cournot duopolists, Shapiro and Stiglitz's (1984) model of efficiency wages, and Barro and Gordon's (1983) model of monetary policy.

[1]Player 1 might wonder whether an opponent who threatens to explode a grenade is crazy. We model such doubts as incomplete information—player 1 is unsure about player 2's payoff function. See Chapter 3.

In Section 2.4 we introduce the tools necessary to analyze a general dynamic game of complete information, whether with perfect or imperfect information. We define the *extensive-form* representation of a game and relate it to the normal-form representation introduced in Chapter 1. We also define subgame-perfect Nash equilibrium for general games. The main point (of both this section and the chapter as a whole) is that a dynamic game of complete information may have many Nash equilibria, but some of these may involve noncredible threats or promises. The subgame-perfect Nash equilibria are those that pass a credibility test.

2.1 Dynamic Games of Complete and Perfect Information

2.1.A Theory: Backwards Induction

The grenade game is a member of the following class of simple games of complete and perfect information:

1. Player 1 chooses an action a_1 from the feasible set A_1.
2. Player 2 observes a_1 and then chooses an action a_2 from the feasible set A_2.
3. Payoffs are $u_1(a_1, a_2)$ and $u_2(a_1, a_2)$.

Many economic problems fit this description.[2] Two examples

[2]Player 2's feasible set of actions, A_2, could be allowed to depend on player 1's action, a_1. Such dependence could be denoted by $A_2(a_1)$ or could be incorporated into player 2's payoff function, by setting $u_2(a_1, a_2) = -\infty$ for values of a_2 that are not feasible for a given a_1. Some moves by player 1 could even end the game, without player 2 getting a move; for such values of a_1, the set of feasible actions $A_2(a_1)$ contains only one element, so player 2 has no choice to make.

(discussed later in detail) are Stackelberg's model of duopoly and Leontief's model of wages and employment in a unionized firm. Other economic problems can be modeled by allowing for a longer sequence of actions, either by adding more players or by allowing players to move more than once. (Rubinstein's bargaining game, discussed in Section 2.1.D, is an example of the latter.) The key features of a dynamic game of complete and perfect information are that (i) the moves occur in sequence, (ii) all previous moves are observed before the next move is chosen, and (iii) the players' payoffs from each feasible combination of moves are common knowledge.

We solve a game from this class by backwards induction, as follows. When player 2 gets the move at the second stage of the game, he or she will face the following problem, given the action a_1 previously chosen by player 1:

$$\max_{a_2 \in A_2} u_2(a_1, a_2).$$

Assume that for each a_1 in A_1, player 2's optimization problem has a unique solution, denoted by $R_2(a_1)$. This is player 2's *reaction* (or best response) to player 1's action. Since player 1 can solve 2's problem as well as 2 can, player 1 should anticipate player 2's reaction to each action a_1 that 1 might take, so 1's problem at the first stage amounts to

$$\max_{a_1 \in A_1} u_1(a_1, R_2(a_1)).$$

Assume that this optimization problem for player 1 also has a unique solution, denoted by a_1^*. We will call $(a_1^*, R_2(a_1^*))$ the *backwards-induction outcome* of this game. The backwards-induction outcome does not involve noncredible threats: player 1 anticipates that player 2 will respond optimally to *any* action a_1 that 1 might choose, by playing $R_2(a_1)$; player 1 gives no credence to threats by player 2 to respond in ways that will not be in 2's self-interest when the second stage arrives.

Recall that in Chapter 1 we used the normal-form representation to study static games of complete information, and we focused on the notion of Nash equilibrium as a solution concept for such games. In this section's discussion of dynamic games, however, we have made no mention of either the normal-form representation or Nash equilibrium. Instead, we have given a

verbal description of a game in (1)–(3), and we have defined the backwards-induction outcome as the solution to that game. In Section 2.4.A we will see that the verbal description in (1)–(3) is the extensive-form representation of the game. We will relate the extensive- and normal-form representations, but we will find that for dynamic games the extensive-form representation is often more convenient. In Section 2.4.B we will define subgame-perfect Nash equilibrium: a Nash equilibrium is subgame-perfect if it does not involve a noncredible threat, in a sense to be made precise. We will find that there may be multiple Nash equilibria in a game from the class defined by (1)–(3), but that the only subgame-perfect Nash equilibrium is the equilibrium associated with the backwards-induction outcome. This is an example of the observation in Section 1.1.C that some games have multiple Nash equilibria but have one equilibrium that stands out as the compelling solution to the game.

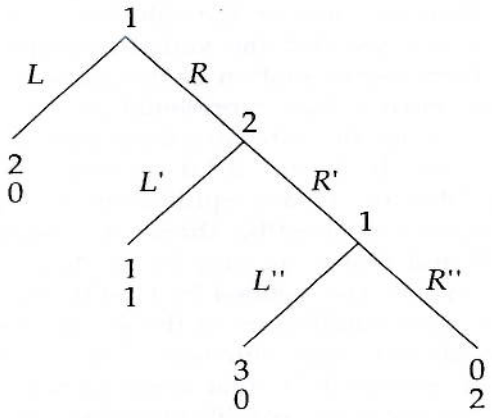
We conclude this section by exploring the rationality assumptions inherent in backwards-induction arguments. Consider the following three-move game, in which player 1 moves twice:

1. Player 1 chooses L or R, where L ends the game with payoffs of 2 to player 1 and 0 to player 2.

2. Player 2 observes 1's choice. If 1 chose R then 2 chooses L' or R', where L' ends the game with payoffs of 1 to both players.

3. Player 1 observes 2's choice (and recalls his or her own choice in the first stage). If the earlier choices were R and R' then 1 chooses L'' or R'', both of which end the game, L'' with payoffs of 3 to player 1 and 0 to player 2 and R'' with analogous payoffs of 0 and 2.

All these words can be translated into the following succinct game tree. (This is the extensive-form representation of the game, to be defined more generally in Section 2.4.) The top payoff in the pair of payoffs at the end of each branch of the game tree is player 1's, the bottom player 2's.



To compute the backwards-induction outcome of this game, we begin at the third stage (i.e., player 1's second move). Here player 1 faces a choice between a payoff of 3 from L'' and a payoff of 0 from R'', so L'' is optimal. Thus, at the second stage, player 2 anticipates that if the game reaches the third stage then 1 will play L'', which would yield a payoff of 0 for player 2. The second-stage choice for player 2 therefore is between a payoff of 1 from L' and a payoff of 0 from R', so L' is optimal. Thus, at the first stage, player 1 anticipates that if the game reaches the second stage then 2 will play L', which would yield a payoff of 1 for player 1. The first-stage choice for player 1 therefore is between a payoff of 2 from L and a payoff of 1 from R, so L is optimal.

This argument establishes that the backwards-induction outcome of this game is for player 1 to choose L in the first stage, thereby ending the game. Even though backwards induction predicts that the game will end in the first stage, an important part of the argument concerns what would happen if the game did not end in the first stage. In the second stage, for example, when player 2 anticipates that if the game reaches the third stage then 1 will play L'', 2 is assuming that 1 is rational. This assumption may seem inconsistent with the fact that 2 gets to move in the second stage only if 1 deviates from the backwards-induction outcome of the game. That is, it may seem that if 1 plays R in the first stage then 2 cannot assume in the second stage that 1 is rational, but this is not the case: if 1 plays R in the first stage then it cannot be common knowledge that both players are rational, but there

remain reasons for 1 to have chosen R that do not contradict 2's assumption that 1 is rational.[3] One possibility is that it is common knowledge that player 1 is rational but not that player 2 is rational: if 1 thinks that 2 might not be rational, then 1 might choose R in the first stage, hoping that 2 will play R' in the second stage, thereby giving 1 the chance to play L'' in the third stage. Another possibility is that it is common knowledge that player 2 is rational but not that player 1 is rational: if 1 is rational but thinks that 2 thinks that 1 might not be rational, then 1 might choose R in the first stage, hoping that 2 will think that 1 is not rational and so play R' in the hope that 1 will play R'' in the third stage. Backwards induction assumes that 1's choice of R could be explained along these lines. For some games, however, it may be more reasonable to assume that 1 played R because 1 is indeed irrational. In such games, backwards induction loses much of its appeal as a prediction of play, just as Nash equilibrium does in games where game theory does not provide a unique solution and no convention will develop.

2.1.B Stackelberg Model of Duopoly

Stackelberg (1934) proposed a dynamic model of duopoly in which a dominant (or leader) firm moves first and a subordinate (or follower) firm moves second. At some points in the history of the U.S. automobile industry, for example, General Motors has seemed to play such a leadership role. (It is straightforward to extend what follows to allow for more than one following firm, such as Ford, Chrysler, and so on.) Following Stackelberg, we will develop the model under the assumption that the firms choose quantities, as in the Cournot model (where the firms' choices are simultaneous, rather than sequential as here). We leave it as an exercise to develop the analogous sequential-move model in which firms choose prices, as they do (simultaneously) in the Bertrand model.

The timing of the game is as follows: (1) firm 1 chooses a quantity $q_1 \geq 0$; (2) firm 2 observes q_1 and then chooses a quantity

[3] Recall from the discussion of iterated elimination of strictly dominated strategies (in Section 1.1.B) that it is common knowledge that the players are rational if all the players are rational, and all the players know that all the players are rational, and all the players know that all the players know that all the players are rational, and so on, ad infinitum.

$q_2 \geq 0$; (3) the payoff to firm i is given by the profit function

$$\pi_i(q_i, q_j) = q_i[P(Q) - c],$$

where $P(Q) = a - Q$ is the market-clearing price when the aggregate quantity on the market is $Q = q_1 + q_2$, and c is the constant marginal cost of production (fixed costs being zero).

To solve for the backwards-induction outcome of this game, we first compute firm 2's reaction to an arbitrary quantity by firm 1. $R_2(q_1)$ solves

$$\max_{q_2 \geq 0} \pi_2(q_1, q_2) = \max_{q_2 \geq 0} q_2[a - q_1 - q_2 - c],$$

which yields

$$R_2(q_1) = \frac{a - q_1 - c}{2},$$

provided $q_1 < a - c$. The same equation for $R_2(q_1)$ appeared in our analysis of the simultaneous-move Cournot game in Section 1.2.A. The difference is that here $R_2(q_1)$ is truly firm 2's reaction to firm 1's observed quantity, whereas in the Cournot analysis $R_2(q_1)$ is firm 2's best response to a hypothesized quantity to be simultaneously chosen by firm 1.

Since firm 1 can solve firm 2's problem as well as firm 2 can solve it, firm 1 should anticipate that the quantity choice q_1 will be met with the reaction $R_2(q_1)$. Thus, firm 1's problem in the first stage of the game amounts to

$$\begin{aligned} \max_{q_1 \geq 0} \pi_1(q_1, R_2(q_1)) &= \max_{q_1 \geq 0} q_1[a - q_1 - R_2(q_1) - c] \\ &= \max_{q_1 \geq 0} q_1 \frac{a - q_1 - c}{2}, \end{aligned}$$

which yields

$$q_1^* = \frac{a - c}{2} \quad \text{and} \quad R_2(q_1^*) = \frac{a - c}{4}$$

as the backwards-induction outcome of the Stackelberg duopoly game.[4]

[4]Just as "Cournot equilibrium" and "Bertrand equilibrium" typically refer to the Nash equilibria of the Cournot and Bertrand games, references to "Stackelberg equilibrium" often mean that the game is sequential- rather than simultaneous-move. As noted in the previous section, however, sequential-move games sometimes have multiple Nash equilibria, only one of which is associated with the backwards-induction outcome of the game. Thus, "Stackelberg equilibrium" can refer both to the sequential-move nature of the game and to the use of a stronger solution concept than simply Nash equilibrium.

Recall from Chapter 1 that in the Nash equilibrium of the Cournot game each firm produces $(a-c)/3$. Thus, aggregate quantity in the backwards-induction outcome of the Stackelberg game, $3(a-c)/4$, is greater than aggregate quantity in the Nash equilibrium of the Cournot game, $2(a-c)/3$, so the market-clearing price is lower in the Stackelberg game. In the Stackelberg game, however, firm 1 could have chosen its Cournot quantity, $(a-c)/3$, in which case firm 2 would have responded with its Cournot quantity. Thus, in the Stackelberg game, firm 1 could have achieved its Cournot profit level but chose to do otherwise, so firm 1's profit in the Stackelberg game must exceed its profit in the Cournot game. But the market-clearing price is lower in the Stackelberg game, so aggregate profits are lower, so the fact that firm 1 is better off implies that firm 2 is worse off in the Stackelberg than in the Cournot game.

The observation that firm 2 does worse in the Stackelberg than in the Cournot game illustrates an important difference between single- and multi-person decision problems. In single-person decision theory, having more information can never make the decision maker worse off. In game theory, however, having more information (or, more precisely, having it known to the other players that one has more information) *can* make a player worse off.

In the Stackelberg game, the information in question is firm 1's quantity: firm 2 knows q_1, and (as importantly) firm 1 knows that firm 2 knows q_1. To see the effect this information has, consider the modified sequential-move game in which firm 1 chooses q_1, after which firm 2 chooses q_2 but does so without observing q_1. If firm 2 believes that firm 1 has chosen its Stackelberg quantity $q_1^* = (a-c)/2$, then firm 2's best response is again $R_2(q_1^*) = (a-c)/4$. But if firm 1 anticipates that firm 2 will hold this belief and so choose this quantity, then firm 1 prefers to choose its best response to $(a-c)/4$—namely, $3(a-c)/8$—rather than its Stackelberg quantity $(a-c)/2$. Thus, firm 2 should not believe that firm 1 has chosen its Stackelberg quantity. Rather, the unique Nash equilibrium of this

modified sequential-move game is for both firms to choose the quantity $(a-c)/3$—precisely the Nash equilibrium of the Cournot game, where the firms move simultaneously.[5] Thus, having firm 1 know that firm 2 knows q_1 hurts firm 2.

2.1.C Wages and Employment in a Unionized Firm

In Leontief's (1946) model of the relationship between a firm and a monopoly union (i.e., a union that is the monopoly seller of labor to the firm), the union has exclusive control over wages, but the firm has exclusive control over employment. (Similar qualitative conclusions emerge in a more realistic model in which the firm and the union bargain over wages but the firm retains exclusive control over employment.) The union's utility function is $U(w, L)$, where w is the wage the union demands from the firm and L is employment. Assume that $U(w, L)$ increases in both w and L. The firm's profit function is $\pi(w, L) = R(L) - wL$, where $R(L)$ is the revenue the firm can earn if it employs L workers (and makes the associated production and product-market decisions optimally). Assume that $R(L)$ is increasing and concave.

Suppose the timing of the game is: (1) the union makes a wage demand, w; (2) the firm observes (and accepts) w and then chooses employment, L; (3) payoffs are $U(w, L)$ and $\pi(w, L)$. We can say a great deal about the backwards-induction outcome of this game even though we have not assumed specific functional forms for $U(w, L)$ and $R(L)$ and so are not able to solve for this outcome explicitly.

First, we can characterize the firm's best response in stage (2), $L^*(w)$, to an arbitrary wage demand by the union in stage (1), w. Given w, the firm chooses $L^*(w)$ to solve

$$\max_{L \geq 0} \pi(w, L) = \max_{L \geq 0} R(L) - wL,$$

the first-order condition for which is

$$R'(L) - w = 0.$$

[5] This is an example of a claim we made in Section 1.1.A: in a normal-form game the players choose their strategies simultaneously, but this does not imply that the parties necessarily *act* simultaneously; it suffices that each choose his or her action without knowledge of the others' choices. For further discussion of this point, see Section 2.4.A.

Prof. dr. Tine STANOVNIK

JAVNE FINANCE

LJUBLJANA, 2008

Poglavje 1

Vloga in funkcije javnega sektorja

1.1 Uvod

Predmet Ekonomika javnega sektorja je posvečen študiju ekonomskih zakonitosti, načinu financiranja in delovanju javnega sektorja. Formalno je področje javnih financ nekoliko ožje, ker je poudarek na proučevanju le ene plati javnega sektorja, to je prihodkovne strani oziroma financiranja; "formalno" pravimo zato, ker je skoraj nemogoče povsem ločeno obravnavati prihodkovno in odhodkovno stran javnih financ. Javne finance opravljajo tri osnovne funkcije. To so:

a) **Alokacija** produkcijskih tvorcev oziroma finančnih virov. To je proces, s katerim se opravlja ne samo razdelitev resursov na produkcijo javnih in produkcijo zasebnih dobrin, temveč tudi nadaljnja alokacija znotraj skupine javnih dobrin.

b) **Prerazdelitev dohodka**. S prerazdelitvijo dohodka javne finance praviloma popravljajo tržne izide, in to tako, da se prerazporeja dohodek od premožnejših k manj premožnim.

c) **Stabilizacija gospodarstva**. To pomeni, da naj bi država s primernimi inštrumenti fiskalne in monetarne politike zasledovala določene

makroekonomske cilje, kot so nizka brezposelnost, nizka inflacija, visoka gospodarska rast ipd.

Obstaja tudi četrta funkcija, ki dejansko ni samostojna, temveč bi jo lahko uvrstili v alokacijsko funkcijo. To je **regulatorna** funkcija, s katero država zagotavlja in "skrbi" za inštitucije in pravni red, ki so potrebni za nemoteno delovanje zasebnega sektorja.

Regulatorna funkcija praviloma ni deležna posebne pozornosti javnih financ – kar seveda ne pomeni, da ni pomembna. Posledice odsotnosti regulatorne funkcije so v državah bivše Sovjetske zveze še vidne: brez urejenega pravnega reda in inštitucij, ki zagotavljajo lastninske pravice in veljavnost pogodb, ne more biti dobro delujočega tržnega gospodarstva in zasebnega sektorja. Regulatorna funkcija države se kaže tudi v tem, da predpisuje in nalaga zasebnemu sektorju spoštovanje različnih tehničnih norm in standardov, kar vpliva na stroške in učinkovitost zasebnega sektorja. Na kratko prikažimo osnovne značilnosti posameznih funkcij javnih financ.

1.2 Alokacijska funkcija in nepopolnosti trga

Pod določenimi pogoji tržni mehanizem zagotavlja učinkovito alokacijo resursov; ti pogoji so, da imamo na danem trgu mnogo kupcev in mnogo prodajalcev, da je produkt homogen, da mejni stroški produkcije naraščajo, da potrošnja ali proizvodnja ne povzroča zunanjih učinkov in da imamo vsi na voljo popolno informacijo o produktu. Če ti pogoji niso izpolnjeni, tržni mehanizem praviloma ne zagotavlja učinkovite alokacije resursov in takrat govorimo o nepopolnosti trga *(market failure)*. Z drugimi besedami, tržni mehanizem bo v teh primerih proizvajal bodisi preveč bodisi premalo določenih dobrin glede na družbeno želeno raven produkcije. Prikažimo nekaj najpomembnejših vzrokov za nepopolnost trga in nakažimo obliko državnega oziroma javnega posredovanja.

Obstoj javnih dobrin. To so dobrine, pri katerih ni smotrno izključiti posameznika iz potrošnje; mejni stroški oskrbe, tj. stroški oskrbe dodatnega posameznika, so enaki nič. Ker torej velja $MC = 0$, je optimalno cenovno pravilo $P = MC = 0$. Zasebni sektor bi seveda za oskrbo zahteval neničelno

ceno. Cena torej ne bi bila enaka mejnim stroškom oskrbe, zato bi prišlo do izgube družbene blaginje, če bi oskrbo prepustili zasebnemu sektorju.

Zunanji učinki (eksternalije). Mnoge dobrine porajajo bodisi na strani proizvodnje bodisi na strani potrošnje pozitivne ali negativne zunanje učinke. Pri negativnih zunanjih učinkih je proizvodnja (ali potrošnja) določene dobrine višja od družbeno zaželene, ker tržni mehanizem ne more upoštevati dejstva, da potrošnja ali proizvodnja zvišuje stroške ali znižuje raven koristnosti drugim subjektom. Termoelektrarna ne proizvaja samo električne energije, temveč tudi negativne eksternalije v obliki pepela, žveplovega dioksida, ogljikovega dioksida itd.; soseda, ki vzorno skrbi za svoj vrt in ga lahko vsi sprehajalci občudujejo, pa "proizvaja" pozitivne eksternalije. Skratka, eksternalije nastopajo takrat, ko v mojo funkcijo koristnosti ne vstopa kot argument samo "moja" potrošnja dobrin, temveč tudi potrošnja dobrin drugih oseb. V primeru pozitivnih eksternalij se moja koristnost ob večji potrošnji drugih oseb povečuje, v primeru negativnih eksternalij pa se moja koristnost zaradi večje potrošnje drugih oseb (oziroma proizvodnje) zmanjšuje. Tako lepo urejen in viden vrt pri sosedi pozitivno vpliva na mojo funkcijo koristnosti, onesnažen zrak pa vpliva negativno. Država problem zunanjih učinkov rešuje tako, da pri pomembnih eksternalijah obdavčuje proizvodnjo ali potrošnjo ali skuša z mehanizmi regulacije (postavitev emisijskih standardov itd.) zmanjšati te učinke. Pri dobrinah, ki porajajo pozitivne eksternalije, država lahko tudi neposredno posega v oskrbo (npr. zdravstvo, šolstvo).

Nepopolna konkurenca. Nepopolna konkurenca pomeni, da imamo padajoče mejne stroške oziroma rastoče ekonomije obsega; to je primer naravnega monopola. Kot je znano, je v primeru naravnega monopola (tj. pri padajočih mejnih stroških proizvodnje) output, pri katerem monopolist maksimira svoj dobiček (to je točka, kjer je mejni strošek enak mejnemu prihodku), nižji od outputa v primeru popolne konkurence (to je točka, kjer je mejni strošek enak ceni). S tem da se proizvaja nižji output od družbeno želenega, pride do neto izgube družbene blaginje.

Nepopolnost ("asimetrija") informacij. Trg zagotavlja učinkovito alokacijo, če ima, med drugim, kupec na voljo popolno in natančno informacijo o produktu. Ker ni vedno tako, država lahko v določenih primerih prevzema obveščanje kupcev (npr. pri cigaretah) ali pa prevzema posledice, ki izvirajo iz slabe obveščenosti kupcev. Tako je npr. pred leti Stanovanjski sklad Republike

Slovenije ponudil ugodne stanovanjske kredite vsem tistim, ki so zaradi slabe obveščenosti vzeli skrajno neugodne stanovanjske kredite komercialnih bank; z ugodnimi krediti so lahko poplačali neugodne. Nekoliko drugače je intervenirala angleška vlada, ki je *de facto* izničila veliko število zelo neugodnih polic življenjskega zavarovanja, ki so jih "nevedneži" kupili pri angleških zavarovalnicah oziroma zasebnih pokojninskih skladih.

Obstoj asimetrije informacij še zdaleč ne pomeni, da bo država "intervenirala" na vseh tistih trgih, kjer npr. prodajalci razpolagajo z več informacijami kot kupci. Predstavljajmo si, da država intervenira in regulira trg rabljenih avtomobilov! Asimetrija informacij se kaže tudi v znanem problemu principala in agenta, kjer principal praviloma razpolaga z bistveno manj informacijami kot agent. S tem so dane možnosti, da agent zasleduje svoje lastne interese, ne pa interesov principala (tj. lastnika kapitala ali plačnika storitev). Problem asimetrije informacij je še zlasti prisoten v zavarovalništvu, in sicer kot problem t. i. negativne selekcije (*adverse selection)*; to bomo obravnavali v poglavju o socialni varnosti.

Negotovost *(uncertainty)*. Za mnoge oblike tveganja zasebni trgi ne obstajajo oziroma ne opravljajo dobro svojih funkcij. Zasebno zavarovalništvo ni pripravljeno prevzeti oskrbe s produkti, za katere je značilno nedoločeno tveganje, tj. tveganje, ki ga ne moremo kvantificirati. V teh primerih država prevzema nase tveganje; socialno zavarovanje je verjetno najpomembnejša inštitucija, ki dejansko "socializira" tveganja in za katero stoji država.

Kaj je skupna značilnost za vse te oblike nepopolnosti trga? Pri vseh teh oblikah nepopolnosti trga točka ravnovesja, določena kot presečišče krivulj ponudbe in povpraševanja in dosežena s tržnim mehanizmom, ne sovpada z družbeno zaželenim ravnovesjem. Praviloma krivulja ponudbe, to je krivulja mejnih stroškov, ne sovpada s krivuljo ponudbe, ki je določena z mejnimi družbenimi stroški. Tako je npr. pri pozitivnih eksternalijah krivulja, ki odraža mejne družbene stroške, pomaknjena na desno od krivulje mejnih (zasebnih) stroškov; podobno je tudi v primeru nepopolne konkurence, kjer je krivulja mejnih družbenih stroškov prav tako pomaknjena na desno itd.

Sam obstoj neke oblike nepopolnosti trga še ni zadosten razlog za poseganje države oziroma za javno oskrbo ali javno financiranje ali pa pravno regulativo. Predstavljajmo si svet, v katerem bi države skušale tako ali drugače rešiti vsak

problem, ki izvira iz negativnih eksternalij. Še pri tistih eksternalijah, ki so res globalne (onesnaževanje zraka, voda in zemlje), so države zelo inertne pri iskanju zadovoljivih rešitev. Večina eksternalij je sicer bolj lokalnega pomena in njihovo reševanje je prepuščeno pravnemu sistemu.

Spreminja se tudi način poseganja države. Nekoč je država pri naravnih monopolih zagotavljala javno oskrbo, danes pa se v mnogih primerih država zadovolji le s tem, da ustrezno regulira naravne monopole. Sicer pa tudi monopoli "niso več tisto, kar so nekoč bili"; tehnološki razvoj in globalizacija sta napravila svoje in sedaj so ekonomsko in tehnološko učinkovite tudi manjše enote (in tudi zniževanje mejnih stroškov ni več tako izrazito). Poleg tega vse živahnejša čezmejna menjava hitro razbija nekoč pomembne nacionalne naravne monopole, npr. pri proizvodnji električne energije.

Podobno se spreminja tudi vloga države pri odpravljanju negotovosti kot vzroka za nepopolnost trga. Tako so se v mnogih evropskih državah lotili reform sistema socialnega zavarovanja, predvsem javnega pokojninskega socialnega zavarovanja, z zmanjševanjem vloge države pri socializiranju različnih oblik tveganja. To v bistvu pomeni, da se negotovost oziroma tveganje delno prenaša z države na posameznika.

Skupni imenovalec oziroma "pravzrok" nepopolnosti trga so transakcijski stroški, ki vključujejo stroške odločitve, stroške informacije, stroške pogajanja in pravne stroške[1]. Visoki transakcijski stroški onemogočajo učinkovito organizacijo trga oziroma preprečujejo, da bi trg sploh nastal. Tako so ti stroški visoki za organizacijo in delovanje trga javnih dobrin, praviloma so tudi visoki za trg eksternalij. "Praviloma" pravimo zato, ker bi bili lahko pri majhnem številu udeležencev na "trgu" in pri natančno definiranih pravilih transakcijski stroški majhni. Tako npr. v ZDA obstaja živahen trg z dovoljenji za emisijo; s tem se delno rešuje problem negativnih eksternalij, vendar je trg močno reguliran, ker država določa kvoto izdanih dovoljenj.

[1] V tej zvezi naj omenimo Coasov izrek, ki pravi, da je ob ničelnih transakcijskih stroških in ob natančno definiranih lastninskih pravicah možno priti do učinkovite rešitve problema eksternalij.

1.3 Prerazdelitvena funkcija

Ena pomembnih funkcij javnih financ je prerazdelitvena funkcija. To prerazdeljevanje naj bi zagotavljalo neko sprejemljivo porazdelitev dohodka oziroma sprejemljivo porazdelitev potrošnje določenih dobrin. Prerazdeljevanje se opravlja v dveh stopnjah.

1. Premožnejši plačujejo višje davke in prispevke kot manj premožni; pri tem niso mišljeni samo absolutno, temveč tudi relativno višji zneski. Tako je za davek od dohodka fizičnih oseb, po naše dohodnina, značilna progresivna obdavčitev, kar pomeni, da je razmerje med plačanimi davki in dohodkom pred obdavčitvijo pri premožnih višje kot pri manj premožnih. Pri davku na potrošnjo so v mnogih državah v veljavi nižje davčne stopnje za nujne življenjske potrebščine itd.

2. Koristi oziroma prejemki *(benefits)*, ki jih dobijo posamezniki od javnega sektorja, niso v razmerju z vplačanimi sredstvi posameznika. Tako je npr. zdravstvena oskrba v javnem sistemu enaka za vse ne glede na vplačane prispevke ali davke. Osnovno izobraževanje ni samo enako, temveč je celo obvezno za vse državljane. Varovanje imetja in premoženja v okviru javnega sistema (policija) je enako za vse itd. Pri dobrinah, ki jih zagotavlja javni sektor, se praviloma zagotavlja enakost oskrbe, dočim se pri različnih denarnih izplačilih (npr. pokojnine, nadomestila za porodniški dopust) v mnogih državah upošteva tudi obseg sredstev (prispevkov), ki jih je v ta namen vplačal posameznik.

Pri ugotavljanju končnih učinkov prerazdeljevanja je treba upoštevati neto koristi *(net benefits)*, ki jih posamezniki prejmejo. To pomeni, da se upoštevajo tako vplačila posameznika v obliki davkov in prispevkov kot tudi koristi, ki jih posamezniki prejmejo. Izračun neto koristi je zelo kompleksen ter osnovan na mnogih predpostavkah, zlasti kar zadeva porazdelitev koristi. Raziskava, ki sta jo za ZDA opravila Pechman in Okner (1974), kaže, da javnofinančno prerazdeljevanje bistveno ne prispeva k pravičnejši porazdelitvi dohodka. Z drugimi besedami to pomeni, da so neto koristi za večino prebivalcev ZDA, razen za sorazmerno majhno najrevnejšo skupino, enake nič.

1.4 Stabilizacijska funkcija

Stabilizacijsko funkcijo opredelimo kot zavestno politično usmerjanje gospodarstva za doseganje visoke zaposlenosti, stabilnosti cen, zadovoljivega salda tekočega računa plačilne bilance in zadovoljive stopnje gospodarske rasti. Ti cilji se ne dosegajo avtomatično, temveč le z usklajenim delovanjem fiskalne in monetarne politike. Žal je sposobnost države, da dosega vse zgoraj omenjene cilje, v današnjih časih zelo omejena. Privlačnost keynesianskega pristopa, ki zagovarja aktivno fiskalno politiko oziroma aktivno poseganje države na makroekonomsko področje, je močno zbledela in stabilizacijska funkcija postaja v svojem delovanju vse bolj omejena. Ni dvoma, da je pomemben razlog za takšno stanje tudi v Maastrichtskih kriterijih (1992) ter kasnejšem Paktu o rasti in stabilnosti (1997) *(Growth and Stability Pact)*, ki na fiskalnem področju postavljata kar precejšnje omejitve članicam Evropske monetarne unije (*European Monetary Union*, EMU).

Poglavje 2

Alokacija javnih dobrin

2.1 Taksonomija dobrin

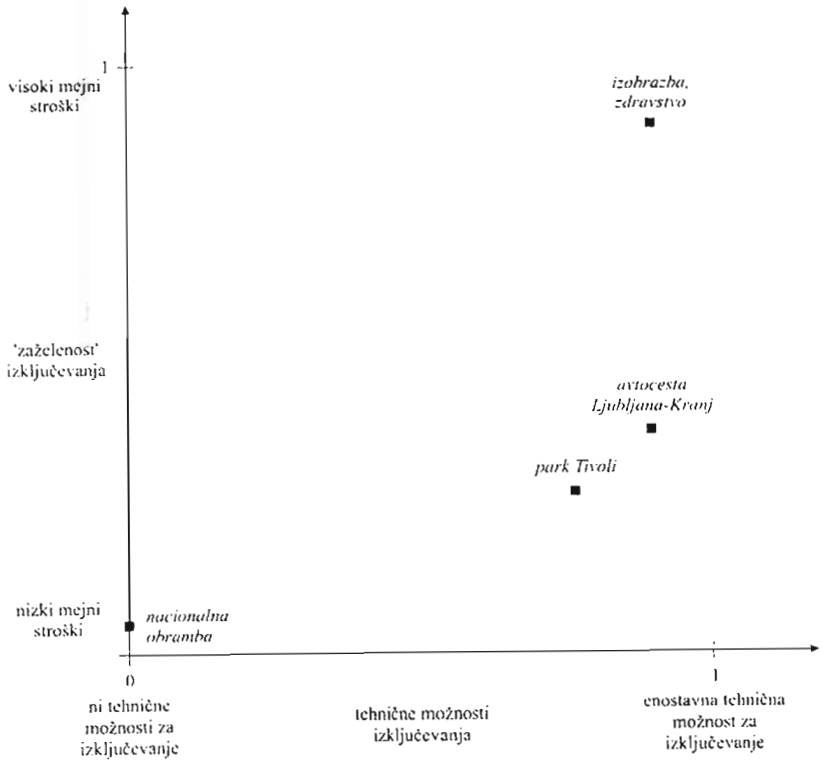
V razdelku 1.2 smo opisali nekaj pomembnih vzrokov za nepopolnost trga; mednje smo prišteli tudi javne dobrine. Javne dobrine smo definirali kot dobrine, pri katerih izključevanje iz potrošnje družbeno ni zaželeno; mejni stroški oskrbe dodatne osebe so enaki nič in bi torej izključitev osebe iz potrošnje imela za posledico povsem nepotrebno izgubo blaginje. Z drugimi besedami: sprememba neto blaginje bi bila negativna.

Za boljše razumevanje pojma javnih dobrin je koristno, da tvorimo določeno taksonomijo dobrin glede na:

a) tehnične možnosti izključevanja iz potrošnje;
b) nekonkurenčnost v potrošnji *(nonrival in consumption)*, tj. v kolikšni meri moja potrošnja dobrine zmanjšuje razpoložljivost te dobrine za druge osebe.

Lastnost nekonkurenčnosti je tesno povezana s konceptom mejnih stroškov: če moja potrošnja dobrine ne vpliva na razpoložljivost potrošnje te dobrine pri drugih osebah, to preprosto pomeni, da so mejni stroški oskrbe dodatne osebe enaki nič.

Oglejmo si položaj nekaterih dobrin na osi tehnične izključljivosti in osi "zaželenosti" izključevanja.



Slika 2.1: Taksonomija dobrin

Čiste javne dobrine *(pure public goods)* so dobrine, pri katerih ni možnosti za tehnično izključitev posameznika iz potrošnje, obenem pa izključitev ni zaželena (ker so mejni stroški oskrbe dodatne osebe enaki nič). V tej skupini ni ravno veliko dobrin, mednje uvrščamo predvsem nacionalno obrambo in zagotavljanje javnega reda in miru.

Javne dobrine *(public goods, social goods)* so tiste dobrine, za katere je izključevanje tehnično sicer možno, ni pa zaželeno, ker so mejni stroški oskrbe dodatne osebe enaki nič. Lahko bi sicer postavili ograjo okoli parka Tivoli ali

okoli Blejskega jezera in omejili dostop s pobiranjem vstopnine, vendar bi takšna rešitev pomenila neto izgubo blaginje. Seveda, v obeh primerih predpostavljamo, da se ob prostem dostopu ne pojavlja takšna gneča sprehajalcev, ki bi imela za posledico neničelne mejne stroške oskrbe dodatnega sprehajalca. Položaj dane dobrine na sliki 2.1 torej ni enolično določen, temveč je odvisen od zasedenosti oziroma obiskanosti; prazen Tivoli je lociran blizu abscisne osi, zelo obiskan Tivoli pa nekoliko višje. Ob tem moramo poudariti, da je označitev javnih dobrin neodvisna od strukture lastninskih pravic; čeprav je obala Ženevskega jezera tako rekoč vsa v zasebni lasti, to v ničemer ne spreminja njenega značaja javne dobrine.

Zasebne dobrine *(zasebe goods)* so dobrine, pri katerih je izključevanje tehnično povsem enostavno, obenem pa tudi zaželeno, ker imamo neničelne stroške oskrbe dodatnega posameznika.

Mešane dobrine *(mixed goods, collective goods)* predstavljajo zmes javne in zasebne dobrine. Te dobrine, bodisi v potrošnji ali proizvodnji, povzročajo pozitivne ali negativne eksternalije in te eksternalije so ravno "javno-značajski" del te dobrine.

V teoriji javnih financ obstaja posebna skupina dobrin – to so **meritorne dobrine** *(merit goods)* oziroma **dobrine posebnega družbenega pomena**. Potrošnja teh dobrin sicer poraja nekaj eksternalij, zato bi jih lahko razvrstili v skupino mešanih dobrin, čeprav so v resnici te dobrine zelo blizu zasebnim dobrinam. Za oskrbo s temi dobrinami obstaja neki javni interes. Takšen javni oziroma skupinski interes zagotavlja javno financiranje in/ali javno oskrbo na področju zdravstva, šolstva, storitev socialnega zavarovanja; pri slednjem bi intervencijo države lahko zagovarjali s potrebo po odpravi nepopolnosti trga, ki ima obliko nepopolnih informacij in nedoločenega tveganja. Vključitev teh dobrin v javno oskrbo in/ali javno financiranje se lahko pojasni tudi z določenim paternalizmom države; država odloča, kaj je dobro za njene državljane. Še prepričljivejša je teza o specifičnem egalitarizmu[2]; država nasprotuje prevelikim razlikam pri oskrbi s temi dobrinami in zato jih zagotavlja prek javnega sektorja.

[2] O tem bo več govora v 11. poglavju.

Javni sektor se torej ukvarja z oskrbo tako čistih javnih dobrin kot javnih dobrin in meritornih dobrin. Pri oskrbi s čistimi javnimi dobrinami dejansko ni možno organizirati trga, ker tudi tehnično ni možno izključevanje posameznikov iz oskrbe; pri oskrbi z javnimi dobrinami je organizacija trga možna, toda, kot rečeno, bi tržni mehanizem povzročil neto izgubo blaginje. Zdi se, da je organizacija trga še najmanj problematična pri meritornih dobrinah: za te dobrine država misli, da bi bila krivulja povpraševanja preveč na levi, tj. ob dani ceni bi se povpraševalo po količini, ki je nižja od družbeno zaželene količine. To naj bi bil osnovni razlog za intervencionizem tako pri financiranju kot tudi pri oskrbi s temi dobrinami. Velike razlike v obsegu javnega sektorja med državami nastajajo predvsem zaradi različnega obsega oskrbe s meritornimi dobrinami v okviru javnega sektorja. Tako lahko imamo javno, pa tudi zasebno šolstvo, zdravstvo, pokojninsko zavarovanje itd.

Sedaj, ko smo prikazali taksonomijo dobrin, lahko tudi nekoliko bolj določeno opišemo dejavnost javnega sektorja.

1. Javni sektor pridobiva, večinoma prisilno, del dohodka ekonomskih subjektov (predvsem posameznikov).

2. Pridobljeni dohodek se porablja za
   a) transferje dohodka določenim skupinam prebivalstva

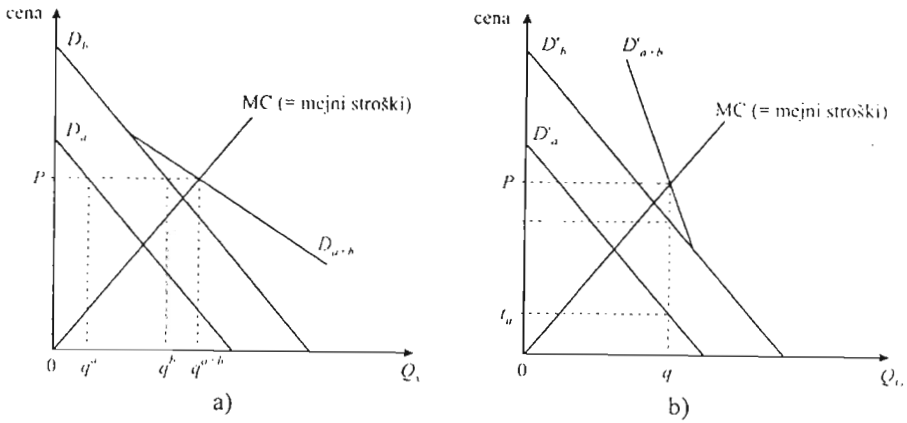
   in za

   b) financiranje ali oskrbo in financiranje določenih dobrin. Te dobrine niso predmet tržne menjave, ker odgovornost za financiranje prevzema javni sektor. Pod "določenimi" dobrinami razumemo čiste javne dobrine, javne dobrine in tudi meritorne dobrine, sam nabor in kvaliteta teh dobrin, ki jih ponuja javni sektor, pa se med državami močno razlikujeta. Tako npr. sta nabor in kvaliteta dobrin, ki jih ponuja javni sektor v skandinavskih državah, večja in boljša od npr. nabora in kvalitete dobrin, s katerimi oskrbuje prebivalstvo javni sektor v ZDA; vse to pa ima svojo ceno: relativni obseg obdavčevanja v Skandinaviji je bistveno višji od obsega obdavčevanja v ZDA.

Iz opisa dejavnosti javnega sektorja vidimo, da točka 2a obravnava prerazdelitveno funkcijo, točka 2b pa alokacijsko funkcijo javnih financ.

2.2 Posebnosti javnih dobrin

Kot nam je dobro znano, se agregatna krivulja povpraševanja za zasebno dobrino dobi tako, da se horizontalno seštevajo individualne krivulje povpraševanja. To izhaja iz čiste konkurenčnosti pri potrošnji teh dobrin: moja potrošnja te dobrine onemogoča potrošnjo iste dobrine (istega jabolka) drugim. Agregatna krivulja povpraševanja po javni dobrini se dobi tako, da se vertikalno seštevajo individualne krivulje povpraševanja; to izvira iz dejstva, da je potrošnja javne dobrine povsem nekonkurenčna, tj. moja potrošnja te dobrine v ničemer ne zmanjšuje dostopa do te dobrine drugim.

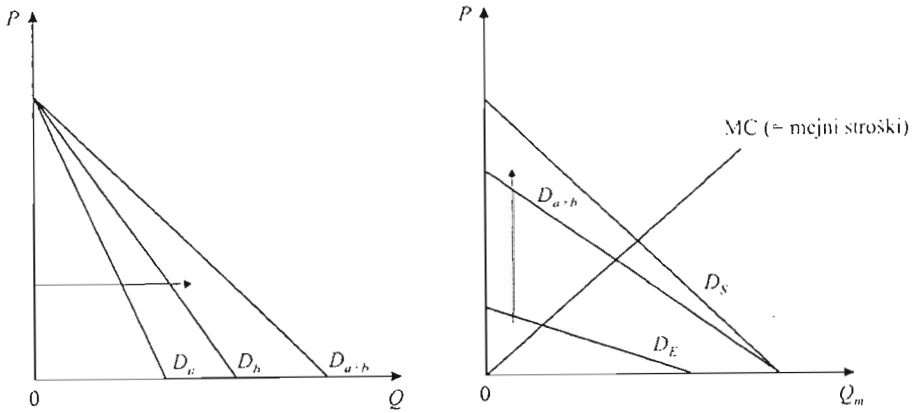


Slika 2.2: Agregatno povpraševanje: primera a) zasebne in b) javne dobrine

V ravnovesju za zasebne dobrine velja: ista cena, različna količina (ki jo trošijo posamezniki), pri javni dobrini pa v točki ravnovesja, tj. v presečišču krivulj ponudbe in povpraševanja, velja: ista količina, različna cena.

Individualne krivulje povpraševanja za mešane dobrine, tj. dobrine, ki imajo značaj tako zasebnih kot javnih dobrin, pa se seštevajo tako, da se zasebni del individualnega povpraševanja sešteva horizontalno, javnoznačajski del individualnega povpraševanja (tj. eksternalije) pa se sešteva vertikalno. Tako je družbeno zaželena ravnovesna cena lahko večja ali manjša od ravnovesne cene, ki je dobljena zgolj s horizontalnim seštevanjem individualnih krivulj povpraševanja; večja je v primeru, ko potrošnja dobrine povzroča pozitivne eksternalije, manjša pa v primeru negativnih eksternalij. Večje eksternalije

poraja potrošnja neke dobrine, večji je njen javnoznačajski del in bolj bo tržno ravnovesje med ponudbo in povpraševanjem odstopalo od družbenega ravnovesja oziroma optimuma. Prikaz določanja družbeno optimalnega ravnovesja pri mešanih dobrinah je podan na sliki 2.3. Kot vidimo, imamo tu kombinacijo horizontalnega seštevanja individualnih krivulj povpraševanja po zasebni dobrini (to je značaj zasebne dobrine) in vertikalno seštevanje individualnih krivulj povpraševanja po javni dobrini (to je značaj javne dobrine). Na levi strani slike 2.3 se individualne krivulje povpraševanja seštevajo horizontalno in tako dobimo agregatno krivuljo povpraševanja D_{a+b}. Na desni strani slike se javnoznačajski del povpraševanja, to je D_E, vertikalno sešteva z dobljeno agregatno krivuljo povpraševanja D_{a+b} in tako dobimo družbeno zaželeno krivuljo povpraševanja D_S. Težava je v tem, da ni tržnega mehanizma, ki bi zagotovil tvorbo krivulje povpraševanja D_S oziroma imamo dejansko le krivuljo povpraševanja D_{a+b} .



Slika 2.3: Krivulja povpraševanja za mešane dobrine

Individualnim krivuljam povpraševanja po javni dobrini rečemo krivulje kvazi-povpraševanja iz preprostega razloga: teh krivulj ni moč izmeriti, kajti pri povpraševanju po javnih dobrinah posameznik nima motiva, da bi razkril svoje prave preference; če izhajamo iz "nealtruističnega" obnašanja, bo posameznik zainteresiran, da plača čim nižjo ceno za oskrbo z javno dobrino.

2.3 Normativna teorija alokacije javnih dobrin

To, da govorimo o krivuljah kvazipovpraševanja po javni dobrini, nas že napeljuje k domnevi, da tržni mehanizem ne more zagotoviti optimalne alokacije javnih dobrin. Zdi se torej, da se mora dejanska alokacija javnih dobrin opraviti z drugimi mehanizmi, ki jih bomo prikazali kasneje, v razdelkih 2.5 in 2.6. V tem razdelku bomo prikazali t. i. normativno teorijo alokacije.

Sama normativna teorija v bistvu izhaja iz osnovnih postulatov mikroekonomije ter konstrukcijsko oziroma s pomočjo formaliziranih izpeljav določa optimalno alokacijo med zasebnimi in javnimi dobrinami; teorija torej z dejanskim, resničnim določanjem alokacije nima nikakršne zveze in zgolj prikazuje, "kako naj bi" potekal postopek alokacije, ter izpeljuje lastnosti točke optimuma.

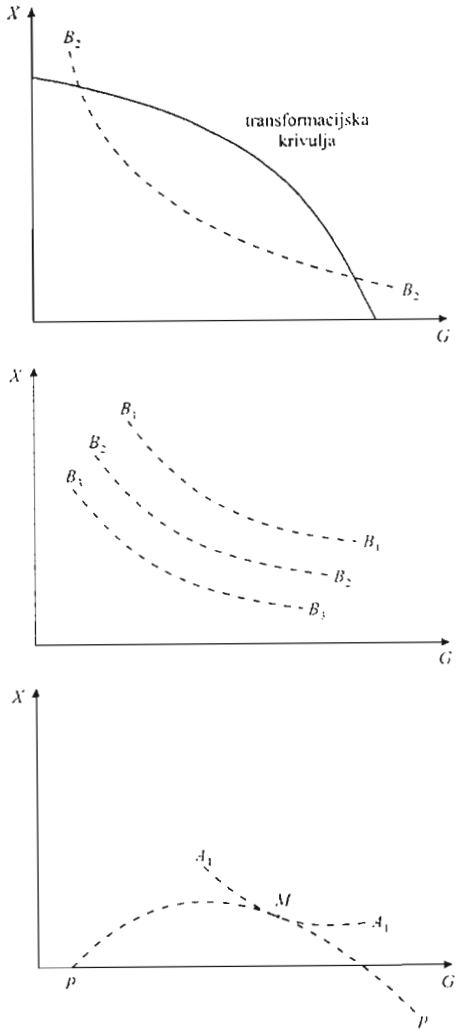
2.3.1 Samuelsonova rešitev

Rešitev problema alokacije javnih dobrin, ki jo je postavil Paul Samuelson, je preprosta in elegantna. Članek, v katerem je prikazal grafično rešitev (Samuelson, 1955), pa je verjetno eden najpogosteje citiranih člankov v ekonomski literaturi.

Imamo osebi A in B ter dve dobrini: X je zasebna dobrina, G pa javna dobrina. Na zgornjem delu slike 2.4 je prikazana dana transformacijska krivulja *(production possibility curve)*, to je krivulja, ki kaže maksimalno možno produkcijo javne dobrine G in zasebne dobrine X ob danih produkcijskih faktorjih. V srednjem delu slike 2.4 je podano polje indiferenčnih krivulj (tj. krivulj konstantne koristnosti) za osebo B. Vzemimo neko indiferenčno krivuljo, naj bo to B_2B_2.

Če od ordinat transformacijske krivulje odštevamo ordinate dane indiferenčne krivulje B_2B_2, dobimo krivuljo možne potrošnje *(consumption possibility curve)* za osebo A. Ta krivulja je podana s PP in se ji včasih reče tudi krivulja ostankov *(leftover curve)*, ker kaže, kakšne kombinacije potrošnje javne dobrine G in zasebne dobrine X so na voljo osebi A ob dejstvu, da je oseba B na izbrani indiferenčni krivulji. Pri dani krivulji možne potrošnje za osebo A poiščemo najvišje ležečo indiferenčno krivuljo osebe A; to je indiferenčna

krivulja, ki je tangentna na krivuljo možne potrošnje. Konstrukcijsko smo dobili točko, ki kaže maksimalno možno dosegljivo koristnost za osebo A ob dejstvu, da je oseba B na dani indiferenčni krivulji, tj. da oseba B dosega dano koristnost. Alokacija je očitno učinkovita po Paretu, ker smo na transformacijski krivulji (tj. krivulji najvišje možne produkcije), oseba A pa je maksimirala svojo koristnost ob dani fiksni koristnosti osebe B.



Slika 2.4: Javne in zasebne dobrine v splošnem ravnovesju

Kakšne so formalne lastnosti te rešitve? Očitno velja, da je ordinata transformacijske krivulje minus ordinata indiferenčne krivulje osebe B (tj. B_2B_2) enako ordinata krivulje možne potrošnje osebe A.

Če diferenciramo zgoraj v besedah opisan izraz in zapišemo

MST = mejna stopnja transformacije,
MSS = mejna stopnja substitucije,

dobimo:

$MST_{XG} = MSS^{B}_{XG}$ + odvod krivulje možne potrošnje osebe A.

V tangentni točki M pa očitno velja:

$$MST_{XG} = MSS^{B}_{XG} + MSS^{A}_{XG}$$

in je to torej pogoj za Paretovo učinkovito alokacijo. Drugače zapisano:

$$MST_{XG} = \sum_{A,B} MSS_{XG} \qquad (2.1)$$

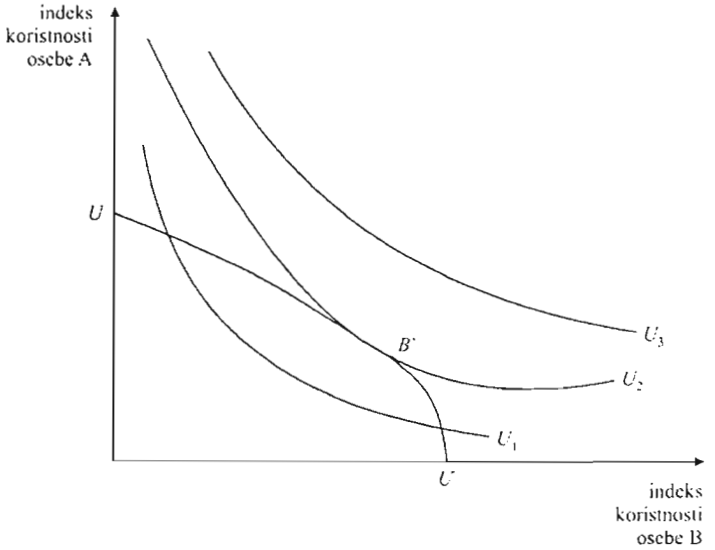
Ta pogoj za Paretov optimum alokacije med javno in zasebno dobrino lahko primerjamo z znanimi pogoji za Paretov optimum v primeru, ko imamo dve zasebni dobrini (npr. X in Y) in dve osebi (npr. A in B). V tem primeru je pogoj za Paretov optimum podan z:

$$MST_{XY} = MSS^{A}_{XY} = MSS^{B}_{XY} \qquad (2.2)$$

Pogoj za Paretov optimum pri alokaciji javnih dobrin, tj. da je seštevek mejnih stopenj substitucije (v potrošnji) enak mejni stopnji transformacije (v proizvodnji), izvira iz dejstva, da je potrošnja javnih dobrin nekonkurenčna oziroma več oseb troši enako količino javne dobrine.

Tako smo konstrukcijsko dobili eno točko Paretovega optimuma in ugotovili, da ta točka zadošča pogoju 2.1. Postopek ponovimo z izbiro neke druge indiferenčne krivulje osebe B (npr. B_3B_3); dobimo novo krivuljo možne

potrošnje za osebo A in zopet novo točko, v kateri je neka indiferenčna krivulja osebe A tangentna na krivuljo možne potrošnje. Tako počasi zbiramo dvojice točk koristnosti za osebi A in B; te točke kažejo maksimalno koristnost osebe A ob dani koristnosti osebe B in na osnovi teh dvojic lahko izrišemo krivuljo meje koristnosti *UU' (utility frontier)*, kot je to razvidno na sliki 2.5.



Slika 2.5: Določitev optimalne točke pri Samuelsonovi konstrukciji

Vse točke na meji koristnosti so učinkovite po Paretu, torej nimamo enolične rešitve problema alokacije. Samuelson ta problem reši tako, da določi družino družbenih indiferenčnih krivulj (funkcije družbene blaginje, *social welfare function*): posamezna krivulja oziroma funkcija določa vse dvojice točk koristnosti za osebi A in B, pri katerih je družba indiferentna in ki jim družba pripisuje enako vrednost. Z družbenega vidika je torej najboljša tista točka, pri kateri se neka družbena indiferenčna krivulja še dotika meje koristnosti; na sliki 2.5 je to točka B^*.

2.3.2 Lindahlova rešitev

Švedski ekonomist Erik Lindahl je leta 1919 postavil "svojo" teorijo alokacije javnih dobrin. Lindahlov model izhaja iz individualnih krivulj povpraševanja

po javni dobrini. Postopek poteka takole: zopet predpostavljamo dve osebi (A in B) ter javno dobrino G in zasebno dobrino X. Zapišimo proračunski omejitvi za A in B:

$$(h_A p_G)\, G + p\, X_A = R_A$$
$$(h_B p_G)\, G + p\, X_B = R_B \qquad (2.3)$$

kjer je h_A delež cene javne dobrine G, ki jo je pripravljena plačati oseba A, h_B pa delež cene javne dobrine G, ki jo je pripravljena plačati oseba B. R_A je dohodek osebe A, R_B pa dohodek osebe B. Očitno mora veljati $h_A + h_B = 1$. Označimo tudi $X_A + X_B = X$ in brez kakršnih koli omejitev na splošno lahko predpostavimo, da je $p = 1$ (to je pač preprosto normiranje). Zapišimo neko indiferenčno krivuljo za osebo A:

$$U^A\ (X_A, G) = \text{konst.} \qquad (2.4)$$

Točke krivulje povpraševanja po javni dobrini dobimo enako kot za zasebno dobrino: pri dani proračunski premici poiščemo indiferenčno krivuljo, ki se dotika proračunske premice. Z drugimi besedami, iščemo maksimalno možno dosegljivo indiferenčno krivuljo za osebo A, ob dani proračunski premici. Ta problem sicer lahko rešimo kot problem vezanega ekstrema, lahko pa se rešitve lotevamo postopoma. Namreč, v točki, kjer se indiferenčna krivulja dotika proračunske premice, morata biti hkrati izpolnjena naslednja pogoja:

$$U^A_{X_A}\, dX_A + U^A_G\, dG = 0 \qquad (2.5)$$

in

$$h_A p_G\, dG + dX_A = 0 \qquad (2.6)$$

torej

$$U^A_{X_A}(-h_A p_G\, dG) + U^A_G\, dG = 0 \qquad (2.7)$$

in

$$MSS^{A}_{XG} = U^{A}_{G} / U^{A}_{X_A} = h_A p_G \text{ oziroma } U^{A}_{G} = h_A p_G U^{A}_{X_A} \quad (2.8a)$$

Podobno dobimo

$$MSS^{B}_{XG} = U^{B}_{G} / U^{B}_{X_B} = h_B p_G \text{ oziroma } U^{B}_{G} = h_B p_G U^{B}_{X_B} \quad (2.8b)$$

Povedano z besedami: za osebo A mora na krivulji povpraševanja za javno dobrino veljati naslednja enakost: mejna koristnost potrošnje javne dobrine je enaka produktu cene, ki bi jo oseba A plačala za javno dobrino, in mejne koristnosti potrošnje zasebne dobrine. Podobno velja tudi za osebo B.

Agregatna krivulja povpraševanja je dana z enačbo (upoštevajoč $h_A + h_B = 1$):

$$\sum_{A,B} MSS_{XG} = p_G(h_A + h_B) = p_G \quad (2.9)$$

Transformacijsko krivuljo lahko zapišemo kot

$$p_G G + X_A + X_B = \text{konst. (upoštevaje } p = 1) \quad (2.10)$$

oziroma

$$p_G G + X = \text{konst.}$$

Vse točke na krivulji ponudbe javne dobrine morajo torej zadoščati pogoju

$$p_G \, dG + dX = 0 \quad (2.11)$$

sledi

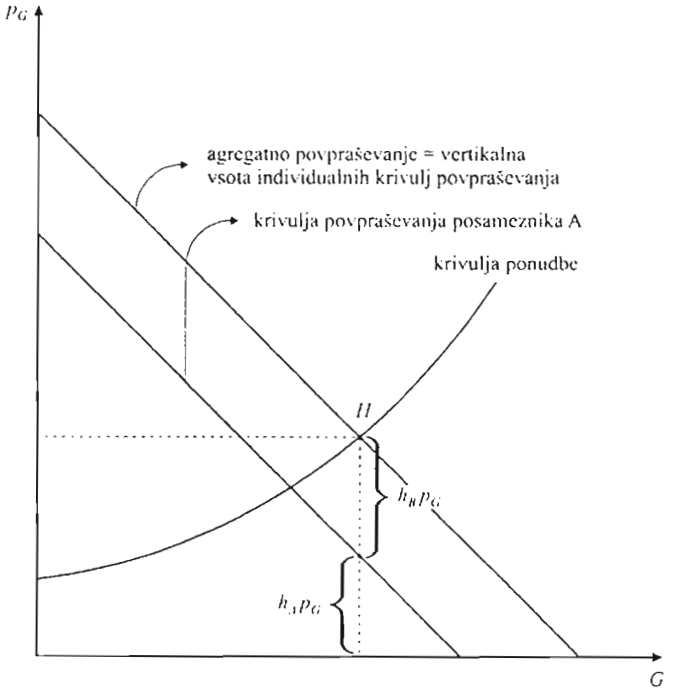
$$MST_{XG} = -\frac{dX}{dG} = p_G \quad (2.12)$$

V točki ravnovesja, tj. v presečišču krivulje agregatnega povpraševanja in krivulje agregatne ponudbe, morata hkrati veljati tako izraz (2.9) kot izraz (2.12). Iz tega sledi, da mora v točki ravnovesja veljati:

$$\sum_{A,B} MSS_{XG} = MST_{XG} \quad (2.13)$$

Z drugimi besedami, točka Lindahlovega ravnovesja je tudi učinkovita po Paretu.

Samo rešitev si lahko grafično ponazorimo na sliki 2.6.



Slika 2.6: Lindahlovo ravnovesje

Jasno je, da bi nekoliko drugačna porazdelitev dohodka rezultirala v drugačnih individualnih krivuljah povpraševanja in torej tudi v drugačni agregatni krivulji povpraševanja ter končno v drugačni točki ravnovesja. Pri Lindahlovi konstrukciji tako kot pri Samuelsonovi obstaja neštetо točk Paretovega optimuma. Pri premikanju iz ene točke Paretovega optimuma v drugo pa seveda eden od posameznikov izgubi, drugi pa dobi. Tudi če se fiksira dohodek obeh udeležencev in postavi pravilo obdavčevanja (oziroma "plačevanja

davčne cene" za javno dobrino G), ni mehanizma, ki bi zagotavljal doseganje ravnovesne točke. Zaradi nekonkurenčnosti potrošnje javne dobrine posamezniki niso zainteresirani, da bi razkrili svoje preference; kot smo že omenili, individualnim krivuljam povpraševanja po javnih dobrinah prav zaradi nezmožnosti, razkriti dejanske preference posameznikov, rečemo krivulje kvazipovpraševanja. Dodatni problem v razkrivanju preferenc je v tem, da bo tisti, ki bi ob dani davčni ceni povpraševal po večji količini javne dobrine G, v ravnovesju plačal tudi višjo davčno ceno P_G. Ravno zaradi tega bo v njegovem interesu, da svojih pravih preferenc ne razkrije. Tega "skrivanja preferenc" seveda pri zasebnih dobrinah ni, ker v ravnovesju vsi plačajo isto ceno.

2.3.3 Primerjava med obema rešitvama

Obe konstrukcijski metodi, tj. oba postopka normativne alokacije javnih dobrin, nas pripeljeta do enakega pogoja Paretove učinkovite alokacije javnih in zasebnih dobrin. Ta pogoj je:

$$\sum_{A,B} MSS_{XG} = MST_{XG}$$

Poti, ki nas pripeljeta do tega pogoja, pa sta različni. Oba postopka neposredno upoštevata produkcijsko učinkovitost, kar pomeni, da se ves čas nahajamo na transformacijski krivulji oziroma na meji produkcijskih možnosti. Samuelson na osnovi indiferenčnih krivulj izpelje pogoje za učinkovitost po Paretu, Lindahl pa na osnovi krivulj kvazipovpraševanja po javni dobrini. Pri obeh konstrukcijah enolična rešitev ni rezultat nekih tržnih mehanizmov; pri Samuelsonu je določitev ene same rešitve problema omogočena predvsem z "*deus ex machina*", tj. z določitvijo oziroma postavitvijo funkcije družbene blaginje, pri Lindahlu pa je odvisno od "sporazumevanja" obeh skupin. Zaradi odsotnosti tržnega mehanizma za doseganje ravnovesne točke tudi razvrščamo obe metodi v skupino normativnih postopkov alokacije javnih dobrin.

2.4 Javna izbira

2.4.1 Arrowov izrek nemožnosti

Končni rezultat normativne teorije alokacije je negativen. To pomeni, da oba postopka, Samuelsonov in Lindahlov, pripeljeta sicer do optimalne alokacije, tj. do točke Paretovega optimuma, vendar ne razkrivata mehanizma, ki bi zagotavljal, da se to ravnovesje tudi doseže. Pri Samuelsonu je optimum, tj. "najboljši" Paretov optimum, dobljen s pomočjo funkcije družbene blaginje, pri Lindahlu pa je optimum odvisen od pogajalskih moči posameznih udeležencev; če so udeleženci enakovredni (tj. imajo enako pogajalsko moč), potem naj bi bili pripravljeni razkriti svoje preference in tudi s pogajanji doseči optimum. Nadalje, pri Samuelsonu je "najboljši" optimum odvisen od same funkcije družbene blaginje, pri Lindahlu pa od porazdelitve dohodka; drugačna funkcija družbene blaginje (pri Samuelsonu) oziroma drugačna porazdelitev dohodka (pri Lindahlu) bi imela za posledico neko drugo točko, ki pa bi bila prav tako optimalna po Paretu.

Ob študiju Samuelsonove konstrukcije in rešitve na osnovi funkcije družbene blaginje se zastavlja vprašanje, ali bi bilo možno iz individualnih preferenc na osnovi določenega pravila dobiti družbeno preferenco. Temu pravilu rečemo pravilo družbene izbire. Izkaže se, da v splošnem ni takega pravila, s katerim bi z agregacijo individualnih preferenc dobili družbeno preferenco, ne da bi pri tem to pravilo kršilo vsaj enega od petih sicer smiselnih pogojev, ki jih bomo navedli v nadaljevanju. Ta izjemno pomembna ugotovitev je podana v Arrowem izreku nemožnosti (*Arrow's impossibility theorem*). V besedah avtorja izreka:
"Če izključimo možnost primerjav koristnosti med osebami, bodo za širok razpon individualnih razvrstitev definirane edino tiste metode prehoda od individualnih preferenc k družbenim preferencam, ki so bodisi vsiljene bodisi diktatorske." (Arrow, 1994, str. 68)

Arrowov izrek nemožnosti se pretežno prikazuje na primeru individualnega glasovanja o posameznih izbirah.

Kakšni so torej pogoji, ki bi jih po Arrowu moralo izpolnjevati pravilo, ki agregira individualne preference in tvori neko skupno preferenco oziroma razvrstitev? Navedimo osnovne pogoje:

1. **Pogoj racionalnosti**. Za vsako množico individualnih preferenc mora pravilo družbene izbire dati razvrstitev, ki je popolna in tranzitivna. Popolnost je definirana tako, da je vsaka razvrstitev primerljiva z vsako drugo (tj. pri primerjavi dveh razvstitev smo bodisi indiferentni ali preferiramo eno razvrstitev); tranzitivnost pa pomeni: če imamo tri stanja ("alternative") in če A preferiramo B, B pa preferiramo C, potem tudi A preferiramo C. Pogoj racionalnosti dejansko pomeni, da zahtevamo od pravila družbene izbire tisto, kar zahtevamo tudi od pravila individualne izbire (tj. funkcije koristnosti). Predstavljajmo si npr. pravilo individualne izbire (tj. funkcije koristnosti), ki ne bi imelo lastnost tranzitivnosti!
2. **Pogoj neodvisnosti od irelevantnih alternativ**. Družbena razvrstitev alternativ A in B je odvisna samo od tega, kako posamezniki razvrščajo A in B. Družbena razvrstitev alternativ A in B ni odvisna od tega, kako posamezniki razvrščajo A in B glede na neko tretjo alternativo C. Na družbeno izbiro torej ne smejo vplivati alternative, ki niso "v igri", tj. alternative, ki niso relevantne[3].
3. **Paretovo načelo**. Če vsak posameznik preferira alternativo A alternativi B, potem bo tudi pravilo družbene izbire tako, da bo alternativa A preferirana alternativi B. Če vsaj en posameznik preferira alternativo A alternativi B, vsi ostali pa so indiferentni, potem mora biti pravilo družbene izbire tako, da je alternativa A preferirana alternativi B.
4. **Pogoj neomejene domene**. Pravilo družbene izbire mora biti takšno, da upošteva vse možne individualne razvrstitve (preference). Z drugimi besedami, ne želimo takšne družbene izbire, ki bi omejevala ali izključevala posameznika, pod pogojem, da so individualne razvrstitve posameznika racionalne.
5. **Pogoj nediktatorstva**. Ne obstaja posameznik, čigar preference so samodejno tudi preference družbe. To pomeni, da če bi za vsako možno dvojico alternativ A in B neki posameznik ("diktator") preferiral alternativo A alternativi B, potem bi družba preferirala alternativo A alternativi B ne glede na individualne preference drugih posameznikov.

[3] Pojem alternativa uporabljamo tukaj širše: ne samo kot izbor med dvema možnostma, temveč tudi izbor med več možnostmi.

Pogoji, ki jih postavlja nobelovec Kenneth Arrow za pravilo družbene izbire (oziroma funkcijo družbene blaginje), so sicer navidezno sprejemljivi, a vendarle preveč omejevalni. Kot bomo videli kasneje, klasično večinsko glasovanje ne izpolnjuje vseh Arrowovih pogojev, ker ne izpolnjuje pogoja racionalnosti. Z drugimi besedami, družbeno razvrščanje, ki sledi iz tega pravila družbene izbire (tj. pravilo: večinsko odločanje), ni tranzitivno. Res je, da takšen nepričakovan rezultat nastane zaradi nekoliko neobičajnih oblik individualnih preferenc[4]. Tudi zaradi tega kritiki Arrowega izreka nemožnosti poudarjajo, da so pogoji, ki jih mora izpolniti funkcija družbene blaginje oziroma pravilo družbene izbire, nekoliko preostri.

Po drugi strani pa lahko Arrowove pogoje pojmujemo kot referenčno točko, primerno za primerjavo posameznih funkcij družbene blaginje, tj. pravil, ki iz individualnih preferenc tvorijo neko skupno družbeno razvrščanje oziroma izbiro. V tem smislu se Arrowovi pogoji lahko primerjajo s pogoji za obstoj popolnega trga; tudi tu imamo referenčno točko, ki nam služi za primerjavo (z vidika družbene blaginje) posameznih oblik trgov.

2.4.2 Volilna pravila

Kot smo torej ugotovili, je prevedba individualnih preferenc v neko družbeno izbiro problematična, ker pravila družbene izbire – v splošnem – ne zadoščajo hkrati vsem petim Arrowovim pogojem. To je seveda teoretična ugotovitev. Dejstvo je namreč, da se vsakodnevno določajo družbene izbire, in to praviloma tako, da se individualne preference "prevedejo" v neko družbeno funkcijo blaginje oziroma družbeno izbiro z glasovanjem.

V politični sferi, kamor spadajo tudi odločitve o alokaciji javnih dobrin, se pri glasovanju praviloma uporablja načelo "en človek, en glas". Kot pravi Musgrave (1993, str. 94), se demokracija v glavnem pojmuje tako, da se "kombinira radikalno egalitaristično načelo 'en človek, en glas' z neegalitaristično delitvijo 'dolarskih glasovnic' v ekonomski sferi".
Vendar pravilo "en človek, en glas" ne zadošča za definiranje volilnega pravila (tj. volilne procedure). Volilno pravilo mora namreč vsebovati način določanja zmagovite alternative. Eno zanimivih vprašanj je, pri katerih volilnih pravilih in kakšnih dodatnih pogojih lahko vendarle pridemo do "dobre" družbene

[4] To dejansko pomeni, da vse individualne preference niso z enim samim maksimumom.

izbire. May (1952) je pokazal, da je v primeru, ko imamo dve alternativi, večinsko odločanje nedvomno najboljše volilno pravilo. Pri določanju volilnega pravila v primeru treh alternativ se stvari precej zapletejo. Kot smo že omenili, pri takšnem naboru alternativ pravilo večinskega odločanja lahko krši pogoj racionalnosti (natančneje: pogoj tranzitivnosti). Nadalje, v primeru, da imamo več kot dve alternativi, je navadno večinsko odločanje podvrženo t. i. manipulaciji agende. To pomeni, da prisotnost ali odsotnost določenih alternativ lahko vpliva na izbor zmagovite alternative.

Edino možno "zdravilo" za manipulacijo agende je takšno pravilo družbene izbire, ki je sicer osnovano na večinskem odločanju, a na primerjavah vseh parov alternativ. Če imamo npr. tri alternative, potem primerjamo A in B, B in C ter C in A. Alternativa, ki zmaga v vseh takšnih primerjavah, je t. i. Condorcetova zmagovalka. Oglejmo si to na primeru, ki kaže na razvrstitev preferenc med 60 volilci.

Tabela 2.1: Paradoks glasovanja

	60 volilcev				
	23	17	2	10	8
Prva preferenca	A	B	B	C	C
Druga preferenca	B	C	A	A	B
Tretja preferenca	C	A	C	B	A

Vir: Boroah, V. K. (1993), str. 140.

Če paroma primerjamo te tri alternative, ugotovimo, da A premaga B (33 glasov proti 27), B premaga C (42 proti 18), toda C premaga A (35 proti 25). Tukaj ne dobimo Condorcetovega zmagovalca, temveč Condorcetov paradoks: čeprav na ravni vsakega posameznika velja tranzitivnost, pa dano pravilo družbene izbire, tj. "večinsko odločanje, paroma primerjanje", ni tranzitivno, tj. ne velja 1. Arrowov pogoj.

Oglejmo si sedaj primer, ko je kršen drugi Arrowov pogoj, tj. pogoj neodvisnosti od irelevantnih alternativ. Tako kot pri prvem pogoju je uporaba tega pogoja smiselna le, ko imamo več kot dve alternativi. Ta pogoj od pravila družbene izbire zahteva, da je družbena razvrstitev alternativ A in B odvisna

samo od tega, kako posamezniki razvrščajo A in B. Če več posameznikov razvršča A > B (tj. alternativa A je pred alternativo B, oziroma alternativa A je nad alternativo B, oziroma alternativa A je višje razvrščena kot alternativa B), potem mora biti tudi družbena razvrstitev taka, da je A > B. Predpostavimo, da imamo glasovalno pravilo (tj. pravilo družbene izbire), ki je takole definirano: večkrožno glasovanje, pri čemer je možno glasovati samo za eno alternativo. Po vsakem glasovanju se izloči alternativa, ki dobi najmanj glasov. Takšno pravilo glasovanja ("pravilo družbene izbire") se uporablja pri volitvi dekana Ekonomske fakultete v Ljubljani ali pri določitvi mesta, ki bo gostilo olimpijado. Tabela 2.2 kaže rezultate glasovanja za določitev olimpijskega mesta za leto 2000.

Tabela 2.2: Kako so glasovali člani Mednarodnega olimpijskega komiteja leta 1993

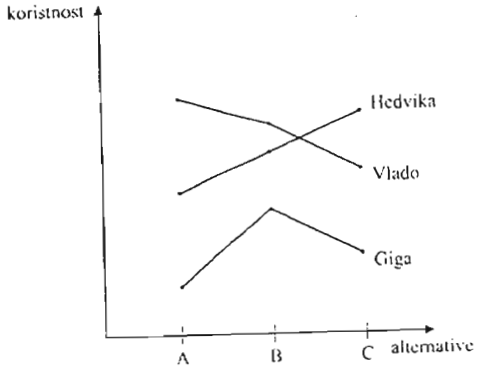
Krog	1	2	3	4
Sydney	30	30	37	45
Peking	32	37	40	43
Manchester	11	13	11	-
Berlin	9	9	-	-
Istanbul	7	-	-	-

Vir: Cowell (2005).

Iz tabele razberemo, da 45 članov MOK razvršča Sydney pred Peking, tj. za 45 članov velja Sydney > Peking, 43 članov MOK pa razvršča Peking pred Sydney, tj. za 43 članov velja Peking > Sydney. Ob tem ni pomembno, na katero mesto člani postavljajo Sydney in Peking: če ima član MOK razvrstitev Berlin > Manchester > Sydney > Istanbul > Peking, v primerjavi Sydneyja in Pekinga to pomeni Sydney > Peking. Torej: če več članov MOK razvršča Sydney pred Peking, bi tudi pravilo družbene izbire (če naj velja pogoj neodvisnosti od irelevantnih alternativ!) moralo biti takšno, da razvrsti Sydney pred Peking. Iz tabele je razvidno, da v prvem, drugem in tretjem krogu pravilo družbene izbire postavlja Peking pred Sydney, kar pomeni, da to pravilo družbene izbire krši pogoj neodvisnosti od irelevantnih alternativ.

Kršitev prvega Arrowovega pogoja ima za posledico ciklično rešitev. Zastavimo si lahko vprašanje, pod kakšnimi pogoji ne pride do cikličnih

rešitev. Izkaže se, da cikličnost nastane zaradi "nenavadne" strukture individualnih preferenc. Predpostavimo, da lahko vse alternative smiselno razvrstimo v eno dimenzijo in da ima vsak volilec najbolj preferirano (tj. "optimalno") alternativo, ki je seveda različna za različne volilce. Nadalje zahtevamo, da se za vsakega volilca preference alternativ desno in levo od optimalne alternative dosledno manjšajo; v tem primeru rečemo, da imajo volilci preference z enim samim maksimumom (*single peaked preferences*). Takšne so npr. preference volilcev na sliki 2.7; za Gigo je npr. maksimum dosežen pri alternativi B, za Vlada je maksimum dosežen pri alternativi A itd.[5]



Slika 2.7: Preference z enim samim vrhom

Predpostavimo torej, da imajo vsi volilci preference z enim samim vrhom (maksimumom). To je možno, če obstaja neka smiselna razvrstitev alternativ v enodimenzionalnem prostoru – npr. alternative odražajo neko politično usmeritev (ekstremno desno, desno, sredinsko, levo, ekstremno levo) ali npr. alternative odražajo stopnjo naklonjenosti za javne izdatke. V primeru, da imajo vsi volilci takšne "lepe" oblike preferenc, potem velja izrek o medianskem volilcu. Ta izrek pove, da bo v vseh medsebojnih primerjavah zmagala tista alternativa, ki jo najbolj preferira medianski volilec, to je tisti volilec, za katerega velja, da ima polovica volilcev preferenčni vrh (tj. maksimum) desno od njegovega maksimuma, polovica volilcev pa ima preferenčni vrh (tj. maksimum) levo od njegovega maksimuma. Iz slike 2.7 je razvidno, da je medianski volilec Giga. Njena preferirana alternativa B bo zmagala v vsaki primerjavi dveh alternativ. Tako bo v primerjavi alternativ B

[5] Matematiki bi rekli, da je lokalni maksimum hkrati tudi globalni.

in A za alternativo B glasovala tudi Hedvika – torej bo alternativa B zmagala z 2 : 1. Pri primerjavi alternativ B in C bo zopet zmagala alternativa B, ker bo zanjo tokrat glasoval (poleg Gige) tudi Vlado.

Glede večinskega odločanja lahko rečemo še to, da takšno odločanje očitno ne zagotavlja učinkovite rešitve (po Paretu), razen v trivialnem primeru, ko imajo vsi posamezniki enako razvrstitev preferenc. V vseh drugih primerih bo po glasovanju prišlo do pozitivnih in negativnih sprememb koristnosti. Tisti, ki bodo utrpeli izgubo, bodo v primeru večinskega odločanja sicer v manjšini in bodo morali prenašati "tiranijo" večine. Ena od slabosti večinskega odločanja je, da ni učinkovita po Paretu[6]; prej smo videli, da pravilo večinskega odločanja tudi ni tranzitivno, to pomeni, da je pri tem pravilu kršen pogoj racionalnosti. Vprašamo se torej, ali obstajajo volilna pravila, ki so učinkovita po Paretu; soglasnost je očitno takšno volilno pravilo, ker bo pri uporabi tega pravila vsaj ena oseba na boljšem in nihče na slabšem. Težava s soglasnostjo je v tem, da povzroča zelo visoke stroške odločitvenega procesa; ti vključujejo poleg stroškov pogajanja tudi oportunitetne stroške, ki nastanejo zaradi odlašanja pri sprejemanju odločitev. V primeru soglasja se torej lahko tiranija večine sprevrže v tiranijo manjšine, in kot pravi Baumol (1965, str. 43), "pravilo soglasja je idealni inštrument za ohranjanje eksternalij in neenakosti, ki že obstajajo".

Zdi se, da ne obstaja neki idealen volilni sistem, ki bi bil hkrati stroškovno učinkovit in učinkovit po Paretu. Seveda ob tem lahko ugotovimo, da pri volilnih sistemih učinkovitost po Paretu ni ravno zaželena lastnost, ker bi upoštevanje te zahteve vnaprej onemogočalo kakršno koli prerazdelitev dohodka.

2.4.3 Referendum o volilnem sistemu v Sloveniji

Volilna pravila oziroma volilni sistem je izjemno pomembna sestavina demokratičnega procesa, lahko bi rekli, da je eden od stebrov demokracije. Tudi Slovenija se je znašla pred preizkušnjo demokracije, ko je poskus spremembe volilnega sistema konec leta 1996 povzročil eno najhujših politično-ustavnih kriz v samostojni Sloveniji. Tako je jeseni 1996 stranka SDS

[6] Da ne bo pomote: to še zdaleč ne pomeni, da je kršen tretji Arrowov pogoj, tj. Paretovo načelo. Večinsko odločanje ni učinkovito po Paretu, ker se ob takšnem odločanju lahko izboljša položaj ene skupine (tj. večine) ob hkratnem poslabšanju položaja druge skupine (tj. manjšine).

sprožila pobudo za spremembo volilnega sistema; predlagala je večinski sistem in v ta namen je zbrala podpise 43.710 volilk in volilcev v podporo referendumu. Ob tem se je zganil državni svet in pristavil svoj "piskrček" s predlogom referenduma o kombiniranem volilnem sistemu. Še istega dne je še stranka LDS podala svoj predlog o proporcionalnem sistemu in v ta namen zbrala podpise 30 poslank in poslancev[7]. Posledica tega referendumskega "tekmovanja" je bila ta, da je državni zbor jeseni 1996 razpisal en sam referendum s tremi referendumskimi vprašanji. V posebnem zakonu o načinu glasovanja in ugotavljanja izida referenduma (Uradni list št. 57, 1996) je državni zbor določil, da je izglasovan tisti predlog (alternativa) "za katerega je glasovala večina volilcev, ki so glasovali". Za dodatno "komplikacijo" je bilo poskrbljeno s tem, da je volilno pravilo omogočalo volilcu, da obkroži "za" pri enem referendumskem vprašanju, lahko pa je (to ni bilo obvezno) obkrožil "proti" pri enem ali obeh (ali celo vseh treh) referendumskih vprašanjih.

Postavljeno volilno pravilo ne zagotavlja nujno tudi zmagovalca; le v primeru, da bi eden od treh predlogov dobil absolutno večino, bi bila ta zmagovalna alternativa tudi Condorcetova zmagovalka. Tudi pravilni postopek izbiranja zmagovalne alternative (s primerjavo parov) nam ne zagotavlja, da bomo dobili Condorcetovo zmagovalko. Ali je bila manipulacija agende ("dodajanje" alternativ) skupaj z volilnim pravilom zavestna odločitev poslank in poslancev? Kaže, da je najmočnejša stranka vladajoče koalicije (LDS) zavestno šla v poskus manipulacije agende.

Po pričakovanju nobeden od ponujenih treh predlogov ni dobil absolutne večine volilcev, ki so glasovali. Ustavno sodišče je takšen "patpoložaj" rešilo tako, da je konec leta 1998 (Uradni list št. 82, 1998) izdalo odločbo, s katero je dalo novo interpretacijo volilnega pravila, določilo zmagovalca referenduma o volilnem sistemu ter naložilo državnemu zboru primerno ukrepanje v roku šestih mesecev. Ustavno sodišče se je očitno dobro zavedalo problema manipulacije agende. Tako v obrazložitvi odločbe navaja: "Nevarnost cepitve glasov in s tem odmika referendumskega izida od dejanske volje volilcev narašča s številom in vsebinsko podobnostjo referendumskih vprašanj. Ureditev, ki jo je prinesla novela zakona o referendumu in ljudski iniciativi, motivira zagovornike veljavne zakonske ureditve, da vlagajo zahteve za razpis

[7] Ciril Ribičič (1998, str. 43) to duhovito opisuje: "... potem, ko je državni svet izglasoval zahtevo za razpis referenduma in je bil dr. Kristan na zasluženem kosilu, so poslanci pod Anderličevo taktirko zbrali podpise in ga prehiteli z vložitvijo svoje zahteve samo zato, da bi preprečili referendum o kombiniranem sistemu."

referenduma samo zato, da bi povzročili cepitev glasov in tako preprečili sprejem kakršne koli odločitve na referendumu." To, kar preseneča, je dejstvo, da je Ustavno sodišče RS skušalo problem manipulacije agende rešiti z manipulacijo volilnega pravila[8]. Argument za tovrstno manipulacijo je bil, da je prvotno sprejeto volilno pravilo možno razumeti na različne načine, zato je določilo, da je izglasovan predlog, "za katerega je glasovala večina volilcev, ki so glasovali pri tem referendumskem vprašanju". Kot volilno pravilo *ex ante* je to nesmiselno, ker omogoča, da so lahko izglasovani prav vsi trije predlogi hkrati. Kot volilno pravilo *ex post* pa je seveda bilo smiselno, ker je le en predlog izpolnil te pogoje, kot je razvidno iz tabele 2.3.

Tabela 2.3: Izid glasovanja referenduma o volilnem sistemu dne 12. decembra 1996

	za	proti	"za" kot % vseh oddanih glasovnic
predlog 1	83,864	237,041	14.4 %
predlog 2	259,687	139,384	44,5 %
predlog 3	152,784	207,965	26,2 %
skupno število oddanih glasovnic: 583,297			
skupno število veljavnih glasovnic 526,390			

Vir: Uradni list RS št. 82, stran 6904.

Opomba: predlog 1 = predlog državnega sveta
predlog 2 = predlog 43.710 volilk in volilcev
predlog 3 = predlog 30 poslank in poslancev

To seveda ni edini greh novopostavljenega volilnega pravila. To volilno pravilo krši načelo "en človek, en glas", ker je s tem, da je prvotno volilno pravilo dopuščalo možnost glasovanja proti, dejansko upoštevalo različno intenziteto preferenc posameznih volilcev. Volilci, ki so zgolj obkrožili "za" pri enem referendumskem vprašanju, so očitno imeli nižjo intenziteto preferenc od volilcev, ki so hkrati z obkrožitvijo "za" pri enem referendumskem vprašanju obkrožili "proti" pri drugih dveh. Skratka, s spremembo volilnega pravila po opravljenem referendumu je Ustavno sodišče RS določilo "zmagovalca" in naložilo državnemu zboru, da v roku šestih mesecev sprejme zakon, ki bo

[8] Z besedami Cirila Ribičiča (1998, str.34): "... odločba Ustavnega sodišča RS je presegla vse dosedanje poskuse, kako presenetiti in spodnesti nasprotnika."

volilni sistem uredil tako, "kot to narekuje izid referenduma"[9]. Gre seveda za izjemno hudo blamažo te najvišje inštitucije sodne veje oblasti; blamažo, ki je skorajda povzročila ustavno krizo v Sloveniji. K sreči je državni zbor poleti 2000 (Uradni list št. 66, 2000) sprejel ustavni zakon, s katerim je nekoliko spremenil dotedanji volilni sistem in preprečil ustavno krizo ter odstranil grožnjo nelegitimnih parlamentarnih volitev[10].

Že iz tega, sicer kratkega in selektivnega prikaza je razvidno, da se raziskovalno področje teorije in prakse družbene izbire osredotoča na analizo načinov agregacije individualnih preferenc ter analizo volilnih pravil, tj. pravil družbene izbire. Analiza posameznih načinov agregacije (ne nujno "en človek, en glas") ter samih volilnih pravil (večinski sistem, sistem kvalificirane večine itd.) nam omogoča primerjavo le-teh glede na postavljeno referenčno točko, to so Arrowovi pogoji.

2.5 Pozitivna teorija alokacije; predstavniška demokracija in interes državne administracije

Dosedanji prikaz problema alokacije javnih dobrin je bil predvsem normativne narave; teorija normativne alokacije javnih dobrin se je ukvarjala s formalnimi postopki za doseganje točke Paretove učinkovitosti in pri tem smo analizirali Samuelsonovo in Lindahlovo rešitev. Arrowov izrek nemožnosti in analiza volilnih postopkov pa sta po svojem značaju prav tako normativna, ker se postavljajo norme (tj. pogoji), ki jih morajo izpolniti "želena" pravila družbene izbire.

Normativna teorija z realnostjo nima veliko zveze. Predvsem velja, da se volilci le redko (razen v Švici) neposredno odločajo o javnih zadevah, pri čemer je to neposredno odločanje v obliki referenduma, kjer imamo praviloma dve alternativi. Volilci praviloma odločajo posredno, prek svojih voljenih predstavnikov – poslancev. Pri tem moramo upoštevati, da poslanci niso

[9] Za odločbo Ustavnega sodišča RS so glasovali sodniki: Miroslava Geč-Korošec, Peter Jambrek, Tone Jerovšek, Lovro Šturm ter Boštjan M. Župančič. Proti so glasovali sodniki: Dragica Wedam Lukić, Franc Testen in Matevž Krivic (Uradni list RS, 82/98).

[10] Spremembe volilnega sistema, ki jih je sprejel državni zbor leta 2000, so podrobno opisane v Ribičič (2003, str. 35–38).

samostojni, temveč pripadniki strank; kot pojasnjujeta Brown in Jackson (1990, str. 110):
"Politične stranke so sestavljene iz skupin posameznikov s skupno ideologijo in skušajo doseči položaj oziroma izvolitev zaradi koristi, ki temu pritičejo (denar, oblast, status). Ideologija je 'splošni pogled na svet'; v negotovem in kompleksnem svetu volilci ne morejo biti popolnoma informirani, ker je to predrago. Ideologija oziroma zavezanost ideologiji zmanjšuje informacijske stroške."

Stranke so torej "zbirno mesto" individualnih preferenc, vendar tudi precej več kot samo "zbiralci glasov". Za stranke je značilna tudi aktivna politika na strani ponudbe, s katero skušajo pridobiti čim več glasov. Tako je po Anthonyju Downu (1957) cilj strank dejansko maksimiranje števila glasov, ne pa učinkovita družbena alokacija resursov; ta je možna le, če konkurenca drugih strank sili vlado (ali vladajočo koalicijo) v premikanje proti točki, ki je učinkovita po Paretu. Takšno gledanje na politični proces je analogno interpretaciji tržnega procesa, kjer proizvajalci med seboj tekmujejo in spremljajo želje potrošnikov.

Na politični proces torej lahko gledamo tudi z optiko ekonomske analize; obstoj strank omogoča volilcem manjše informacijske stroške. Podobno vlogo na trgu proizvodov imajo tržne znamke, ki ob tem, da ponujajo dovolj informacij za nakup, zmanjšujejo informacijske stroške kupca. Pri tržnih znamkah (Lacoste, Nike, Kappa, Benetton) imamo precejšnjo lojalnost kupcev, pri strankah pa precejšnjo lojalnost volilcev. Prodajalci oziroma ponudniki na trgu blaga tako kot politiki na trgu volilnih glasov niso omejeni zgolj na pasivno spremljanje preferenc in maksimiziranje svoje ciljne funkcije (dohodka ali števila glasov), temveč tudi aktivno vplivajo na preference.

Na alokacijo javnih dobrin ne vplivajo samo politiki – poslanci v zakonodajnem telesu; močan vpliv ima tudi birokracija oziroma državni uslužbenci. Kaj skuša maksimizirati ta skupina? Enega od odgovorov je podal W. A. Niskanen (1971, str. 38): Uradniki skušajo maksimizirati velikost svojega ministrstva, agencije ali državnega sklada; s tem si posredno povečujejo ne samo dohodek, temveč tudi status, oblast in vpliv. Uradnik tekmuje z drugimi uradniki za sredstva, podobno kot tekmujejo prodajalci med seboj. Po drugi strani pa ima državni uradnik pomembno prednost pred poslanci: kot ugotavlja Stiglitz (1988, str. 207), je to informacija. Poslanci za

razliko od državnih uradnikov nimajo predstave o dejanskih stroških posameznih javnih storitev. Glede na taktične cilje lahko uradniki zavestno manipulirajo z oceno potrebnih stroškov, tako da jih bodisi precenijo ali podcenijo.

Formalno sicer politiki (tj. ministri) podajajo zahtevek za določitev sredstev iz proračuna za svoje ministrstvo, vendar je ta zahtevek odvisen predvsem od informacij in sugestij uradnikov; odnos med uradnikom in ministrom je v bistvu odnos med agentom in principalom. Tudi odnos med poslanci in volilci bi lahko označili kot odnos med agentom in principalom; odnos izvira iz asimetrije informacij oziroma še bolje: odnos izvira iz neenakih sposobnosti oceniti in ovrednotiti informacijo[11].

2.6 Pozitivna teorija alokacije: kako nastaja državni proračun

Analiza alokacije bi bila nepopolna, če bi se ustavili zgolj pri (normativni ali pozitivni) analizi volilnih pravil, tj. pravil, ki transformirajo individualne preference v skupinsko preferenco, "poosebljeno" v poslancu (oziroma stranki). Izbor poslanca oziroma zastopnika predstavlja le eno fazo v celotnem odločitvenem procesu, kajti osnovna funkcija poslancev, njihov "*raison dêtre*", je odločanje o državnem proračunu. Pri tem opravilu aktivno sodelujejo poleg poslancev (tj. zakonodajnega telesa) tudi vlada (tj. izvršilno telo) in uradniki.

Zgodovinsko gledano je zakonodajno telo sprva odločalo le o davkih; izvršilna oblast (oziroma suveren) ni mogla uvesti davkov, ne da bi bile podane zakonske podlage s strani parlamenta. S tem se je varovala državljanska pravica tistega dela prebivalstva, ki je sploh imel možnost, da izbira predstavnike v parlament, to je pravica, da posredno, prek izvoljenih predstavnikov odloča o obsegu "ekspropriacije", to je, da odloča o davkih[12].

[11] Ni lepšega primera asimetrije informacij, tj. odnosa agenta in principala, kot prikaz zgod in nezgod v angleški TV-nadaljevanki "*Yes, minister*", kjer državni sekretar spretno, s prikrivanjem in doziranjem informacij, vpliva na ministrove odločitve.

[12] Med zgodovinarji ni niti najmanjšega dvoma, da je problem obdavčevanja celo sprožil sam ameriški boj za neodvisnost. Tako so Angleži leta 1764 naložili prebivalstvu ameriških kolonij carino na sladkor, ki naj bi se bolj striktno pobirala kot dotedaj, leta 1765 pa obdavčevanje papirja za tiskanje (to je bilo možno, ker je papir prihajal iz Anglije). Ti davki so sprožili vihar protestov; ameriški priseljenci so načelno nasprotovali takšnemu

Vendar zastopniki oziroma poslanci ne odločajo samo o davkih, tj. o prihodkih, temveč hkrati tudi o odhodkih javnih financ; z drugimi besedami, odločajo o državnem proračunu. In tako kot se individualne preference volilcev "izgubijo" na relaciji med volilcem in poslancem, tako se tudi "izgubijo" individualne preference poslancev ob postopku odločanja in sprejemanja proračuna. Kot bomo videli, je v mnogih državah že sam proračunski postopek takšen, da skuša izvršilna oblast v čim večji meri omejiti svobodo oziroma "prostost" odločanja; razlog za takšno omejevanje gre nedvomno iskati v minimiziranju neposrednih in posrednih stroškov sprejemanja proračuna. Brez takšnih omejitev bi usklajevanje individualnih preferenc poslancev oziroma agregacija teh preferenc potekala zelo počasi, proračuni pa bi se sprejemali z veliko zamudo.

Kako sploh nastaja proračun? Za parlamentarne demokracije so značilne tri stopnje proračunskega procesa:

1. vladna stopnja, pri kateri vlada načrtuje letni proračun in predloži osnutek proračuna parlamentu;
2. parlamentarna stopnja, ki se zaključi s sprejetjem proračuna;
3. stopnja izvrševanja, pri kateri lahko pride tudi do določenih sprememb prvotno sprejetega proračuna.

2.6.1 Vladna stopnja

Vladna stopnja v večini evropskih držav traja 8–9 mesecev (v Sloveniji formalno traja 7 mesecev). Alesina in Perotti (1999) definirata dva različna postopka na vladni stopnji. Pri hierarhičnem postopku ima predsednik vlade ali finančni minister velika pooblastila in lahko "preglasuje" posameznega ministra, ko ta predloži vladi proračun svojega ministrstva. Značilna primera hierarhičnega postopka sta Velika Britanija in Francija. Pri t. i. dogovornem

obdavčevanju, ki je po njihovem kršil "naravne pravice" ameriških kolonov. Vodje odpora so bili namreč trdno prepričani, da angleški parlament ne more obdavčevati ljudstva, saj v njem nima svojega zastopstva. Angleži so na to odgovorili, da ni pomembno realno zastopstvo in da so ameriški koloni virtualno zastopani prek vseh izbranih zastopnikov v parlamentu, ki zastopajo interese ne samo svojih volilcev, temveč interese celotnega angleškega imperija. Takšno pojasnilo je bilo, v besedah zgodovinarja Edmunda Morgana, "slepilna neumnost" (Morgan, 1977, str. 18–24).

Brown, C. V. in Jackson, P. M. (1994), *Economics of the public sector* (poglavje 3, 4 in 7), Blackwell, Oxford.

Munger, Michael C. (2001), "Voting", v: Shughart II, William F. in Laura Razzolini (urednika), *The Elgar Companion to Public Choice,* Edward Elgar, Cheltenham, UK.

Musgrave, R. in Musgrave, P. (1993), *Javne financije u teoriji i praksi* (poglavja 4, 5 in 7), Institut za javne financije, Zagreb.

Ribičič, C. (1998), "*Zakonodajalčeva volja*", Pravna praksa, št. 23.

Rosen, H. S. (1999), *Public finance* (poglavja 5, 6 in 7), Irwin McGraw-Hill, New York.

Stiglitz, J. E. (1988), *Economics of the public sector* (poglavje 5, 6 in 7), W. W. Norton, New York.

Poglavje 3

Uvod v obdavčevanje

3.1 Uvod

Odločitve o obsegu državnega proračuna so hkratne odločitve o prihodkovni in odhodkovni strani javnih financ. Praviloma je prihodkovna stran ozko grlo javnih financ; načrtovani prihodki dostikrat ne dohajajo načrtovanih odhodkov. Posledice so javnofinančni primanjkljaji (fiskalni deficiti), ki so v evropskih državah pogostejši kot javnofinančni presežki (fiskalni suficiti). Sam prikaz odhodkovne strani javnih financ bomo podali v poglavju 14, tukaj pa bomo osvetlili nekatere osnovne pojme in elemente prihodkovne strani.

Osnovni prihodki javnega sektorja so (a) davčni prihodki, (b) transferni prihodki (*grants*) in (c) drugi prihodki.

a) **Davčne prihodke** delimo na davke in prispevke, pri čemer je njihova skupna značilnost vsaj načeloma jasna; oboji so oblika prisilnih dajatev. Za razliko od davkov, ki predstavljajo enostranski prisilni odvzem realnega dohodka oziroma kupne moči, pa smo s plačevanjem prispevkov deležni tudi določenega obsega pravic. Res je, da ta obseg pravic ni natančno definiran in se kar sproti spreminja. To ni npr. samo značilnost slovenskega sistema obveznega zdravstvenega ali pokojninskega zavarovanja, temveč

tudi podobnih sistemov v Evropi. Ob tem je razumljivo, da se obseg pravic lahko spreminja le postopoma; poskusi nenadnih velikih sprememb, predvsem v smeri zmanjšanja pravic, bi se kaj kmalu znašli na "zatožni klopi" ustavnega sodišča. Prispevki so torej politično sprejemljivejša oblika davkov predvsem zaradi "obljube" pravic (pokojnine, zdravstvene oskrbe, itd). Mnogi menijo, da so prispevki po svojih makroekonomskih učinkih podobni davkom, vendar to ni povsem točno. Zaradi obljubljenih pravic si država ustvarja implicitni dolg do prebivalstva in obseg tega implicitnega dolga lahko ima povsem realne posledice, npr. na ceno denarja (tj. na obrestno mero).

b) **Transferni prihodki** (*grants*) vključujejo predvsem prihodke od tujih vlad in mednarodnih organizacij. Za Slovenijo so tu pomembna predvsem sredstva, ki jih prejema iz EU oziroma različnih skladov EU – strukturnega sklada, kohezijskega sklada, socialnega sklada, kakor tudi neposredni transferji kmetijskim proizvajalcem.

c) **Drugi prihodki** so v obliki različnih taks, pristojbin, sredstev od prodaje državnega premoženja, prihodkov od prodaje blaga in storitev, dohodkov od premoženja (dividende, obresti, razdelitev dobička, najemnine).

Davčni prihodki so leta 2006 predstavljali 90,4 %, transferni prihodki 4,7 %, drugi prihodki pa 4,9 % vseh javnofinančnih prihodkov v Sloveniji. Skromen delež drugih prihodkov je razumljiv: zaračunavanje cen za proizvode in storitve javnega sektorja bi (kot smo videli v 2. poglavju) povzročalo neto izgubo blaginje, razen seveda pri tistih proizvodih, ki imajo značaj zasebne dobrine (in pri teh se večinoma pobirajo takse in pristojbine).

Ker so davki *de facto* oblika prisilnega odvzema realnega dohodka prebivalstva, je torej na dlani, da mora biti davčni sistem zasnovan na nekaterih temeljnih načelih, ki zagotavljajo urejeno, zanesljivo, pravično in učinkovito pobiranje davkov ter čim manjše vpletanje države v ekonomski proces. Povod za kmečke punte lahko iščemo v nepravičnih davkih, tako kot je povod za francosko revolucijo v povsem neurejenem in nezanesljivem davčnem sistemu; pod zanesljivostjo tukaj razumemo dejstvo, da posameznik vnaprej ve, kakšna bo njegova davčna obveznost. Če torej skušamo ugotoviti, kakšna so osnovna načela, na katerih mora sloneti sodobni davčni sistem, nimamo v mislih tistih načel ali lastnosti, ki so same po sebi umevne: da npr. davčni sistem zagotavlja zanesljivo določanje davčne obveznosti ali da je zagotovljena politična

odgovornost, s tem da parlament odloča o davkih ipd. Zanimajo nas osnovne, najpomembnejše lastnosti davčnega sistema.

3.2 Zaželene lastnosti davčnega sistema

Seznam zaželenih lastnosti oziroma načel, ki jih mora upoštevati sodobni davčni sistem, se med posameznimi avtorji nekoliko razlikuje, vendar obstaja soglasje, da tri lastnosti po svojem pomenu izstopajo. To so:

1. **Davčni sistem mora biti pravičen**. To pomeni, da mora vsak davčni zavezanec prispevati pravičen delež k financiranju države. V razdelku 3.3 bomo podrobneje pojasnili, kaj to pomeni.
2. **Davčni sistem mora biti tak, da čim manj posega v ekonomske odločitve oziroma da povzroča čim manjšo izgubo učinkovitosti**. Čim manj pravimo zato, ker davčni sistem že s tem, da prisilno odvzema realni dohodek, posega v ekonomske odločitve. Kaj natančno razumemo pod "čim manjšimi" ekonomskimi posledicami oziroma čim manjšo izgubo učinkovitosti? Če bi lahko primerjali (hipotetično) dva davčna sistema, ki imata enak davčni donos, tj. enak znesek pobranih davkov, je boljši tisti sistem, pri katerem je output oziroma proizvod družbe večji. Tako npr. različne oblike obdavčevanja dohodkov od dela ali kapitala lahko povzročijo zelo različne učinke na ponudbo dela ali kapitala; družba je po nepotrebnem odmaknjena od meje produkcijskih možnosti (*production possibility curve)*, tj. od transformacijske krivulje. Tukaj ne gre samo za nekakšne akademske razprave: bistvena zniževanja mejnih davčnih stopenj pri dohodnini v preteklem desetletju (glej tabelo 5.2) so v veliki meri posledica ocen, da visoke mejne davčne stopnje negativno vplivajo na ponudbo dela. Analizi vplivov davkov na odločitve tako v proizvodnji kot v potrošnji bomo posvetili celotno 4. poglavje.
3. **Davčni sistem mora biti tak, da so stroški pobiranja davkov, stroški upravljanja in stroški izpolnjevanja davčne obveznosti nizki (v primerjavi s celotno vrednostjo pobranih davkov)**. Torej od davčnega sistema ne zahtevamo samo nizkih stroškov same davčne administracije, temveč tudi nizke stroške (realne ali oportunitetne) samih davčnih zavezancev, to je fizičnih in pravnih oseb. Tako je npr. znano, da so v ZDA stroški izpolnjevanja davčne obveznosti zelo visoki; Blumenthal in Slemrod (1992) sta ocenila, da je za pripravo dohodninske napovedi davčni

zavezanec v ZDA porabil leta 1989 povprečno 27 ur. O kompleksnosti ameriške dohodnine govori tudi podatek, da si polovica davčnih zavezancev poišče pomoč strokovnjakov za pripravo svojih dohodninskih napovedi[13].

Za razliko od ameriške dohodnine je čas, ki ga porabijo slovenski davkoplačevalci za izpolnitev obveznosti za plačilo dohodnine, bistveno krajši. Maja Klun (2004) je npr. ocenila, da so leta 2000 slovenski davkoplačevalci v povprečju porabili 1,7 ure za izpolnitev dohodninske obveznosti.

Lepo bi bilo, če bi bili ti trije cilji oziroma želene lastnosti davčnega sistema medsebojno neodvisne, vendar žal ni tako. Večja pravičnost oziroma izenačenost ima praviloma za posledico nižjo ekonomsko učinkovitost. Nižji administrativni stroški prav tako "povzročajo" večjo nepravičnost in (verjetno) tudi nižjo ekonomsko učinkovitost. Nizki administrativni stroški pri dohodnini v Sloveniji imajo svojo ceno: obdavčen je predvsem en tip dohodka od dela, ki je preprost za davčno spremljanje, to so plače. Dohodki od kapitala (npr. najemnine) ter mešani dohodki od dela in kapitala (to so dohodki samozaposlenih) so v precejšnjem obsegu neobdavčeni. Razlog je v tem, da je veliko lažje in ceneje spremljati in kontrolirati vidne transakcije, ki se opravljajo prek bančnega sistema, kot pa manj vidne transakcije, ki se pogosto opravijo z gotovinskimi izplačili. Takšne nepravičnosti pri obravnavanju posameznih oblik dohodkov pa vplivajo na učinkovito alokacijo produkcijskih faktorjev.

[13] Visoki administrativni stroški in kompleksnost sistema gresta z roko v roki. O "enostavnosti" ameriške dohodnine priča tudi primer, ki ga navajata Slemrod in Bakija (1996, str. 129): Revija *Money* vsako leto pripravi neko zmerno zahtevno, a fiktivno davčno situacijo posameznika in povpraša davčne strokovnjake – svetovalce, kakšna naj bi bila davčna obveznost tega "fiktivnega" posameznika. V letu 1993 je revija dobila odgovore 41 strokovnjakov. Njihovi odgovori o višini davčne obveznosti tega fiktivnega posameznika so se gibali v razponu od 31.846 USD do 74.450 USD; "pravilno" izračunana davčna obveznost pa je znašala 35.643 USD. "Izstavljeni" računi teh strokovnjakov so prav tako izkazovali kar precejšnjo variabilnost: od 375 USD do 3.600 USD!

3.3 Načela pravičnosti

Obstajata dve splošni načeli pravičnosti (izenačenosti); poudariti moramo, da sta to splošni načeli, ki se primerno prevedeta v davčni načeli.

Načelo horizontalne izenačenosti zahteva enako obravnavanje posameznikov, ki so v enakem položaju glede relevantnih značilnosti.

Načelo vertikalne izenačenosti zahteva različno obravnavanje posameznikov, ki so v različnem položaju glede relevantnih značilnosti.

Načelo horizontalne izenačenosti lahko interpretiramo kot odgovor na vprašanje, na osnovi katerih "relevantnih značilnosti" je sploh dovoljena diskriminacija oziroma razlikovanje med posamezniki. Načelo vertikalne izenačenosti pa skuša odgovoriti na to, kako močna diskriminacija je "dovoljena". Očitno je ključnega pomena določitev nabora "relevantnih značilnosti". Tako se npr. za volilno pravico načelo horizontalne izenačenosti glasi: enako obravnavanje posameznikov (tj. en človek, en glas), pri čemer se kot relevantni značilnosti jemljeta le polnoletnost in državljanstvo. V ne tako oddaljenih časih je bila v nekaterih švicarskih kantonih za volilno pravico relevantna značilnost tudi spol; načelo horizontalne izenačenosti se je torej v njihovi lokalni inačici glasilo: enaka volilna pravica za posameznike, ki so enakega spola. To pomeni, da so vsi moški imeli enako volilno pravico, tj. so lahko glasovali; prav tako so vse ženske imele enako "volilno pravico", tj. niso smele glasovati. Dejansko naj bi bilo načelo horizontalne izenačenosti predvsem načelo, ki preprečuje neupravičeno diskriminacijo; z drugimi besedami, nabor relevantnih značilnosti naj bi bil karseda omejen. V praksi ni vedno tako in dogaja se, da se nabor "relevantnih značilnosti" celo širi. Tako je npr. nemško ustavno sodišče leta 2001 razsodilo, da naj delavci, ki imajo otroke, plačujejo manjši socialni prispevek za dolgotrajno nego ostarelih kot delavci, ki nimajo otrok. Z drugimi besedami, pri načelu horizontalne izenačenosti za plačevanje tega prispevka je ustavno sodišče določilo novo relevantno značilnost – to je število otrok! Logika te razsodbe naj bi bila v tem, da je manjša verjetnost, da bodo starši, ki imajo otroke, potrebovali institucionalno nego v poznih letih, ker naj bi zanje v večji meri skrbeli otroci (!?!).

Pojasnimo načelo horizontalne izenačenosti še na enem zgledu. Pri javnih pokojninskih sistemih, ki so osnovani na načelu socialnega zavarovanja, načelo horizontalne izenačenosti pomeni, da se enako obravnavajo zavarovanci, ki so v pokojninski sistem vplačevali enake prispevke in se upokojili ob isti starosti. "Relevantni značilnosti" sta torej (a) vplačani prispevki in (b) starost ob upokojitvi. Javni sistemi naj ne bi kot relevantno značilnost vključevali npr. (c) spola. Če bi to upoštevali, bi npr. ženske kljub enakim vplačanim prispevkom in enaki starosti ob upokojitvi imele nižjo pokojnino kot moški. Slovenski pokojninski sistem je v preteklosti kot relevantno značilnost vključeval spol, in sicer tako, da je uveljavljal "inverzno diskriminacijo": vstopni pogoji in parametri izračuna pokojnine so bili za ženske ugodnejši kot za moške. Novi pokojninski zakon iz leta 1999 je to "inverzno diskriminacijo" precej omilil, tako da so sedaj med moškimi in ženskami le majhne razlike v vstopnih pogojih in parametrih izračuna pokojnine. Sicer pa so prejšnji pokojninski zakoni v Sloveniji "dodatno" kršili načelo horizontalne izenačenosti s tem, da se je v preteklosti pokojnina izračunavala na osnovi najugodnejšega desetletnega povprečja dohodkov zavarovanca, namesto da bi se upoštevala celotna aktivna doba in s tem – posredno – tudi celotni vplačani prispevki zavarovanca. Novi zakon je to obdobje bistveno podaljšal na najugodnejših osemnajst let in s tem zagotovil boljše upoštevanje načela horizontalne izenačenosti. Če namreč pri izračunu pokojnine upoštevamo le "izsek" delovne dobe, postavljamo v ugodnejši položaj predvsem "bele ovratnike". "Modri ovratniki" so na slabšem, ker imajo enakomernejše dohodke skozi celotno delovno dobo.

Pri načelu vertikalne izenačenosti je predvsem pomembno ugotoviti, kako "različno" naj se obravnavajo posamezniki, ki so različni glede relevantnih značilnosti. V tem smislu je uporaba tega načela težavnejša, samo načelo pa je bolj normativno od načela horizontalne izenačenosti.

Poleg splošnih načel obstajata tudi dve davčni načeli, ki jih formuliramo kot:

- **načelo koristi** (*benefit principle*), ki pravi, da je davčne zavezance treba obdavčiti v skladu s koristmi, ki jih imajo od uporabe javnih dobrin;
- **načelo ekonomske sposobnosti** (*ability-to-pay principle*), ki pravi, da je davčne zavezance treba obdavčiti v skladu z njihovo ekonomsko sposobnostjo, tj. ekonomsko zmožnostjo plačila.

Obe davčni načeli lahko "prevedemo" v temeljni načeli horizontalne in vertikalne izenačenosti, in sicer tako, da namesto "posameznik" vstavimo "davčni zavezanec", namesto splošnega izraza "relevantnih značilnosti" pa vstavimo bodisi specifični pojem "koristi od javnih dobrin" bodisi pojem "ekonomska sposobnost oziroma ekonomska zmožnost plačila". Tako se npr. davčno načelo ekonomske sposobnosti v dikciji temeljnih načel glasi:

- **načelo horizontalne izenačenosti**: enako davčno obravnavanje zavezancev, ki imajo enako ekonomsko sposobnost oziroma enako ekonomsko zmožnost plačila davka;
- **načelo vertikalne izenačenosti**: različno davčno obravnavanje zavezancev, ki imajo različno ekonomsko sposobnost oziroma različno ekonomsko zmožnost plačila davka.

Z načelom koristi se ne srečujemo prvič: že Lindahlova konstrukcija je pokazala, kako naj bi plačevali davke (oziroma "davčno" ceno) v skladu s koristmi, ki jih uporabniki imajo od javne dobrine. Načelo koristi je v splošnem nesprejemljivo in neuporabno, ker ne upošteva ene bistvenih funkcij javnih financ, to je prerazdelitvene funkcije. Kljub temu ima to načelo določeno omejeno uporabnost predvsem tam, kjer imajo javne dobrine ali storitve bolj značaj zasebnih dobrin ali storitev: to so npr. različna dovoljenja, plačevanje javnega prevoza. Načelo je uporabno tudi za obdavčevanje dobrin, ki so komplementarne uporabi javnih dobrin, npr. obdavčevanje bencina za financiranje vzdrževanja cest itd. Poleg tega ima to načelo še eno pozitivno lastnost: povezuje prihodkovno in odhodkovno stran javnih financ.

Načelo ekonomske sposobnosti je torej splošno davčno načelo, ker upošteva tudi možnost prerazdeljevanja oziroma ne povezuje prihodkovne in odhodkovne strani javnih financ. Če torej pristajamo na to načelo, ki je izraženo skozi primerne formulacije načela horizontalne izenačenosti in načela vertikalne izenačenosti, se moramo vendarle vprašati, kako bi konkretizirali tako postavljeni načeli. Kot prvo se moramo vprašati, kako bi primerno definirali ekonomsko sposobnost; med praktičnimi kandidati so predvsem dohodek, potrošnja ali premoženje posameznika oziroma kombinacija le-teh, npr. dohodka in premoženja. Odprto za razpravo je, ali so to res najprimernejši kandidati; ekonomska sposobnost naj bi namreč v idealnih razmerah merila zgolj sposobnost posameznika, da pridobiva dohodek, ne pa dejanske realizacije te sposobnosti, tj. tržne izide, ki se kažejo v dohodku, potrošnji itd.

Nedvomno ima merjenje ekonomske sposobnosti prek tržnih izidov svoje pomanjkljivosti, vendar je takšen "redukcionizem" nujen za neko operativno definicijo ekonomske sposobnosti. Kot drugo, pri konkretizaciji načela vertikalne izenačenosti nastopa vprašanje, kako različno naj se obravnavajo davčni zavezanci. John Stuart Mill, znameniti angleški filozof 19. stoletja, je predlagal, da naj bi različno obravnavanje zavezancev povzročilo pri teh enako žrtev oziroma enako izgubo blaginje. To pomeni, da naj bi se različni davčni zavezanci obravnavali tako, da bi utrpeli enako žrtev. Enaka žrtev se je nadalje operacionalizirala z različnimi merami izgube blaginje oziroma koristnosti, npr. enaka absolutna žrtev, enaka relativna žrtev ali pa enaka mejna žrtev (več o tem v dodatku 1). Lahko se pokaže (Samuelson, 1947, str. 227), da če pod enako žrtev razumemo enako absolutno žrtev, tj. enako izgubo koristnosti pri vsakem posamezniku kot posledico obdavčevanja, potem načelo vertikalne izenačenosti implicira progresivno obdavčevanje dohodka natanko takrat, ko je dohodkovna elastičnost mejne koristnosti dohodka večja od ena. Če označimo z MU funkcijo mejne koristnosti dohodka, z Y pa dohodek, to pomeni

$$\left| \frac{dMU(Y)}{MU} / \frac{dY}{Y} \right| > 1$$

Z drugimi besedami, mejna koristnost dohodka mora zelo hitro padati. Zagovarjati tezo, da je progresivno obdavčevanje posledica uporabe načela vertikalne izenačenosti, je seveda povsem "za lase privlečeno", kajti pogoji, ki morajo biti izpolnjeni (dohodkovna elastičnost mejne koristnosti dohodka mora biti za vse posameznike večja od 1), so zelo omejevalni in končno tudi neugotovljivi. Kot bomo videli v 5. poglavju, so bili teoretični argumenti za progresivno obdavčevanje deležni izrazito ostre in sarkastične kritike v petdesetih letih; delno so zaradi teh kritik davčni sistemi začeli opuščati ostro progresivno obdavčevanje dohodka.

Po vsej tej mentalni vadbi smo dejansko tam, kjer smo bili, to je na začetku. Pozitivna ekonomika ne daje napotila, kako naj se operacionalizira načelo vertikalne izenačenosti, tj. kako različno naj se obravnavajo posamezniki, ki so v različnem ekonomskem položaju oziroma imajo različne ekonomske sposobnosti. To, kar pozitivna ekonomika lahko pove, pa je, da se načelo vertikalne izenačenosti ne more obravnavati izolirano, ne da bi se upoštevale posledice, ki jih ima uporaba tega načela na učinkovito alokacijo resursov.

3.4 Osnovni davčni pojmi

Davčna osnova je vrednost ali količina, na katero se aplicira davčna stopnja. To pomeni:

davčna stopnja · davčna osnova = davek oziroma $t X = T$.

Povprečna davčna stopnja:

$$PDS = \frac{\text{celotni plačani davki}}{\text{davčna osnova}} = \frac{T}{X} \text{ ali } \frac{\Sigma T}{\Sigma X} \qquad (3.2)$$

Mejna davčna stopnja:

$$MDS = \frac{\Delta \text{ celotni plačani davki}}{\Delta \text{ davčna osnova}} = \frac{\Delta T}{\Delta X} \qquad (3.3)$$

Davek je proporcionalen, če je povprečna davčna stopnja konstantna in se torej z rastočo davčno osnovo ne spreminja. To zapišemo kot:

$$\Delta\left(\frac{T}{X}\right) = 0 \qquad (3.4a)$$

Davek je progresiven, če z rastočo davčno osnovo narašča tudi povprečna davčna stopnja. To zapišemo kot:

$$\Delta\left(\frac{T}{X}\right) > 0 \qquad (3.4b)$$

Dohodek fizičnih oseb (dohodnina) je v večini držav (tudi v Sloveniji) obdavčen po stopničasti progresiji. To pomeni, da imamo v posameznih intervalskih vrednostih davčne osnove različne mejne davčne stopnje: npr. 10 % v območju od 0 do 100.000 denarnih enot, 15 % v območju od 100.001 do 150.000 denarnih enot, 20 % v območju 150.001 do 200.000 denarnih enot itd. Če je posameznikova davčna osnova enaka 170.000 denarnih enot, bo na prvih 100.000 enot plačal davek po stopnji 10 %, za naslednjih 50.000 enot bo

plačal davek po stopnji 15 %, za 20.000 enot pa bo plačal davek po stopnji 20 %. Njegova davčna obveznost bo torej 0,10 × 100.000 + 0,15 × 50.000 + 0,20 × 20.000 = 21.500 enot. Indirektna progresija velja, če se davek plačuje šele, ko dohodek preseže neki prag, nad pragom pa velja proporcionalna davčna stopnja. Davek pri indirektni progresiji je podan z relacijo: $T = t\,(X - X_0)$, kjer je X_0 dohodkovni prag. Za razliko od stopničaste progresije je za navadno progresijo značilno, da višja davčna stopnja velja za celotno davčno osnovo. V Sloveniji sta davek na izplačane plače ter davek na motorna vozila v obliki navadne progresije. Tako ima npr. davek na motorna vozila 9 različnih davčnih stopenj, najvišja (trinajstodstotna) pa velja za motorna vozila, katerih prodajna vrednost je nad 25 tisoč evrov. To pomeni, da je celotna vrednost dragih avtomobilov obdavčena po tej stopnji, ne le vrednost nad 25 tisoč evrov.

Davek je regresiven, če z rastočo davčno osnovo pada povprečna davčna stopnja. To zapišemo kot:

$$\Delta\left(\frac{T}{X}\right) < 0 \qquad (3.4c)$$

Efektivno davčno stopnjo *(effective tax rate*, dejanska davčna stopnja) je mogoče definirati na različne načine. Običajno se definira kot razmerje med plačanimi davki in najširšo možno davčno osnovo določenega davka. Tako lahko za posameznega davčnega zavezanca za dohodnino zapišemo:

EDS dohodnine = plačana dohodnina / dohodek pred obdavčitvijo

Podobno zapišemo tudi efektivno davčno stopnjo za posameznega zavezanca za davek od dohodka pravnih oseb:

EDS davka od dohodka pravnih oseb = plačan davek od dohodka pravnih oseb / poslovni presežek (tj. prihodki manj odhodki)

Efektivno davčno stopnjo je seveda mogoče izračunati tudi skupno, za vse dohodninske zavezance ali za vse zavezance davka od dohodka pravnih oseb. Tako je npr. efektivna davčna stopnja dohodnine, na ravni narodnega gospodarstva, določena kot:

EDS dohodnine = celotna plačana dohodnina / celotni dohodek pred obdavčitvijo

Efektivna davčna stopnja dohodnine je sicer pomemben podatek za davčne uprave, a za ekonomsko analizo davčne obremenitve je pomembneje ugotoviti, kakšna je efektivna obremenitev dohodkov posameznih produkcijskih tvorcev (dela in kapitala). Tako je npr. EDS dohodkov od dela enaka:

EDS dohodkov od dela = celotni davki od dohodkov od dela / celotni dohodki od dela (pred obdavčitvijo).

Davčne olajšave *(tax reliefs)* so predvsem dveh oblik:

a) odbitek pred obdavčljivo osnovo *(tax allowance, exemption*, tudi *deduction)* in

b) znižanje davka *(tax credit);* pri tej obliki davčne olajšave se zniža znesek že izračunanega davka.

Davčni izdatki *(tax expenditures)* predstavljajo znesek izpada davčnih prihodkov države zaradi dodeljene davčne olajšave.

Davčni zavezanec je fizična ali pravna oseba, ki ji davčni predpisi nalagajo plačilo davka. Pri številnih davkih davčni zavezanec ni istoveten **davčnemu plačniku**, tj. osebi (fizični ali pravni), ki dejansko nakaže davek na žiro račun proračuna. Tako npr. pri sprotnem obdavčevanju dohodkov fizičnih oseb kot plačnik nastopa podjetje ali organizacija, kjer je davčni zavezanec zaposlen. Ob izplačilih plač podjetje odtegne določen znesek davkov. Takšnemu davku rečemo **davek po odbitku** *(withholding tax)*; razlog za sprotno obdavčevanje pa je, da je najprimernejši trenutek za obdavčevanje ob izplačevanju dohodkov, bodisi dohodkov od dela ali kapitala. Takšno sprotno obdavčevanje z davkom po odbitku je praviloma (toda ne vedno) začasno; ob letni davčni prijavi zavezanec opravi poračun celotnih davčnih obveznosti in že plačanih davkov. **Akontacijski davki**, tj. davki, ki predstavljajo le "predplačilo" davka, so večinoma v obliki davka po odbitku. Pri nekaterih dohodkih v Sloveniji, npr. pri dohodkih od kmetijstva (katastrski dohodek) ali dohodkih od zasebne dejavnosti, pa davčni zavezanec sam sproti nakazuje davčni upravi vnaprej določen znesek tega "predplačila". Sprotno nakazovanje davkov je praviloma

bolj sprejemljivo za davčne zavezance, ker prepreči enkratno veliko obremenitev in plačilo porazdeli na nekaj manjših obrokov.

Davčna utaja *(tax evasion)* je nezakonito izogibanje plačilu davka. **Davčno izogibanje** *(tax avoidance)* predstavlja zakonito izogibanje, s katerim davčni zavezanec prepreči zmanjšanje realnega dohodka. V dejanski praksi ločnica med izogibanjem in utajo ni tako preprosta; nekje na tem neizostrenem mejnem področju je prostor za davčno svetovanje. Kot je nekoč duhovito pripomnil Denis Healey, britanski minister za finance, je razlika med davčnim izogibanjem in davčno utajo v debelini zaporniškega zidu[14]. Ko je obdavčitev odvisna od tega, kako se dejansko opravlja neka aktivnost ali transakcija, govorimo o **davčni arbitraži**. Za davčno arbitražo je značilno, da se sicer doseže isti končni cilj, vendar po zelo različnih "davčnih poteh"; jasno je, da posameznik, če le more, izbere tisto možnost, ki je zanj davčno ugodnejša.

Primer 1. Obdavčitev alkoholnih pijač in tobačnih izdelkov je v Franciji in Belgiji bistveno nižja kot v Veliki Britaniji, zato Angleži "oblegajo" mesta ob francoski in belgijski obali, kupujejo te izdelke in jih prosto prenašajo nazaj v Veliko Britanijo. Takšna oblika davčne arbitraže seveda povzroča Veliki Britaniji velik izpad davčnih prihodkov.

Primer 2. Predpostavimo, da babica, ki živi v Sloveniji, želi podariti stanovanje svojemu dragemu in edinemu vnuku. Za takšen prenos lastnine ima vsaj tri možnosti:

a) stanovanje takoj podari vnuku na osnovi darilne pogodbe;
b) v oporoki zapusti stanovanje vnuku;
c) stanovanje "proda" vnuku.

Z davčnega vidika ni razlik med a) in b); v obeh primerih je transakcija obdavčena po zakonu o davku na dediščine in darila[15]. Če je vrednost stanovanja ocenjena na 150 tisoč evrov, bo vnuk plačal 14.900 evrov, tj. skoraj 10 % ocenjene vrednosti stanovanja. Predpostavimo sedaj, da babica "proda" stanovanje (na osnovi kupoprodajne pogodbe) svojemu vnuku. Po Zakonu o davku na promet nepremičnin[16] je davčna stopnja 2 %. "Prodajalka" bo torej

[14] "The difference between tax avoidance and tax evasion is the thickness of the prison wall."

[15] Uradni list Republike Slovenije, 117/2006.

[16] Uradni list Republike Slovenije, 117/2006.

morala nakazati davčni upravi 3 tisoč evrov; davčna uprava ne preverja, ali je denarna transakcija bila dejansko izvršena. Kot vidimo, izbira "davčne poti" vpliva na plačane davke; v tem primeru znaša razlika 11.900 evrov.

Subjektni in objektni davki. Vsak davek ima svoj subjekt, tj. davčnega zavezanca, in objekt, tj. dejstvo, ki državi omogoča predpisovati davek (dohodek, potrošnja, premoženje itd.). Če davek upošteva bolj same lastnosti subjekta (velikost dohodka, starost, zdravstveno stanje, število in sestava vzdrževanih oseb, način pridobivanja dohodka), potem je to subjektni davek. Subjektni davki so npr. dohodnina in davek od dediščin; pri slednjem se npr. upošteva dedni red dediča. Objektni davki so davki na potrošnjo (tj. prometni davki ali davek na dodano vrednost), carine, socialni prispevki, davek na nepremičnine ipd. Ob tem se zastavlja možnost nadaljnje delitve lastnosti subjekta na (a) dohodkovne in premoženjske značilnosti in (b) druge značilnosti subjekta, kot so npr. starost, zdravstveno stanje, sestava družine itd. Takšna nadaljnja delitev je pomembna ne samo pri davkih, temveč tudi pri negativnih davkih, predvsem pri socialnih prejemkih. Tako npr. dohodnine v posameznih državah v neenaki meri upoštevajo same dohodkovne značilnosti ter druge subjektivne značilnosti davčnega zavezanca. Negativni davki v obliki socialnih prejemkov prav tako upoštevajo tako dohodek posameznika (natančneje: pretekle dohodke oziroma vplačane prispevke) kot tudi specifične značilnosti prejemnika (sestava družine oziroma število otrok, invalidnost, bolezen, nosečnost, starost itd.). Pri davkih je praviloma poudarek na točki (a), tj. na dohodkovnih in premoženjskih značilnostih, pri negativnih davkih pa je poudarek na točki (b), tj. na določenih nedohodkovnih značilnostih subjekta.

Neposredni in posredni davki (direktni in indirektni davki). Ta delitev je nekoliko sporna in jo novejši sistem družbenih računov iz leta 1993 opušča. Neposredni davki so davki, za katere se pričakuje, da bo davčni zavezanec neposredno nosil breme davka: to so dohodnine, davki od dohodka pravnih oseb, davki na dediščine. Posredni davki so davki, za katere se pričakuje, da bremena davka ne bo nosil davčni zavezanec, temveč nekdo drug (praviloma končni potrošnik). Tako je npr. pri davku na dodano vrednost davčni zavezanec sicer podjetje, toda breme tega davka vendarle nosi končni kupec; podobno velja tudi za carine in druge selektivne davke na potrošnjo.

3.5 Priporočena literatura

Musgrave, R. in Musgrave, P. (1993), *Javne financije u teoriji i praksi,* (poglavji 12 in 13), Institut za javne financije, Zagreb.

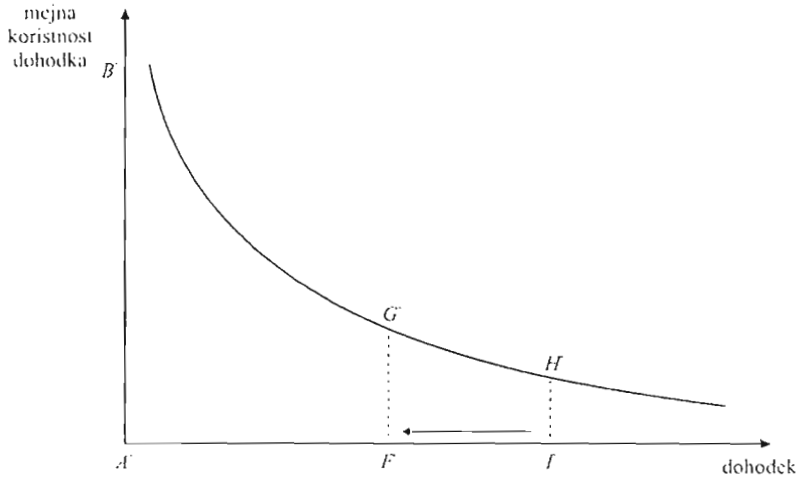
Stiglitz, J. E. (1988), *Economics of the Public Sector,* (poglavje 16), W. W. Norton, New York.

Dodatek poglavju 3

A3.1 Mere enake žrtve

Slika A3.1 kaže funkcijo mejne koristnosti dohodka, pri čemer je ploščina pod krivuljo enaka celotni koristnosti dohodka (za posameznika). Velja namreč:

$$\int \frac{dU}{dY}\, dY = \int dU = U$$



Slika A3.1: Mere enake žrtve

Enaka absolutna žrtev pomeni enako absolutno izgubo koristnosti za vsakega posameznika. To nadalje pomeni, da mora biti ploščina *FGHI* (ki predstavlja izgubo koristnosti ob uvedbi davka) enaka za vse posameznike. Ob tem bi seveda morali upoštevati, da se posamezniki nahajajo na različnih točkah

abscisne osi, tj. da imajo različne dohodke pred obdavčitvijo, kakor tudi da imajo različne krivulje mejne koristnosti dohodka.

Enaka relativna žrtev pomeni enako relativno izgubo koristnosti za vsakega posameznika. Z drugimi besedami, razmerje *FGHI/ABHI* mora biti enako za vse posameznike.

Enaka mejna žrtev pomeni, da mora biti izguba blaginje oziroma koristnosti takšna, da bodo po obdavčitvi vsi posamezniki imeli enako mejno koristnost dohodka. Z drugimi besedami, po obdavčitvi bodo posamezniki imeli takšen dohodek, da bo mejna koristnost tega dohodka (tj. ordinata) enaka za vse.

Zgornji prikaz je ne samo hipotetičen, temveč tudi "stiliziran". Namreč, narisana krivulja mejne koristnosti dohodka predpostavlja fiksne vrednosti drugih parametrov, ki nastopajo kot argumenti v krivulji koristnosti oziroma krivulji mejne koristnosti. To pomeni, da funkcijo koristnosti lahko zapišemo kot U (Y, prosti *čas* = konst.); ob znižanju dohodka (Y), ki nastane kot posledica obdavčevanja, se lahko spremeni količina prostega časa in tako nismo več na isti krivulji mejne koristnosti dohodka.

Poglavje 4

Davčno prevaljevanje, davčna obremenitev in optimalni davki

4.1 Davčno prevaljevanje

Davki so oblika prisilnega odvzema posameznikovega realnega dohodka in zato je razumljivo, da posamezniki skušajo čim bolj zmanjšati posledice takšnega posega. Načini, kako bodo posamezniki reagirali na obdavčevanje, so različni. Tako bodo ob spremembi davka od dohodka nekateri skušali izsiliti povečanje urne mezde oziroma cene dela, spet drugi bodo npr. povečali količino dela, da bi s tem ohranili približno enak realni dohodek kot pred davčno spremembo, spet tretji bodo zmanjšali ponudbo dela v formalnem sektorju gospodarstva itd. Zaradi takšnega različnega prilagajanja posameznikov ne moremo zagotovo trditi, kakšne bodo posledice nekih davčnih sprememb. Z drugimi besedami, ne vemo natančno, kdo bo dejansko utrpel zmanjšanje realnega dohodka zaradi uvedbe davka, tj. kdo bo nosil davčno breme (*tax burden*). Za ugotavljanje davčnega bremena je potrebna analiza procesa prilagajanja ekonomskih subjektov danemu davku, in to od točke, kjer je davek vpeljan, to je od točke zakonske davčne incidence, pa vse do tistih fizičnih oseb, ki se jim je realni dohodek zaradi uvedbe davka dejansko znižal (to predstavlja ekonomsko davčno incidenco). Procesu prilagajanja, ki poteka od točke zakonske davčne incidence pa vse do

Dr. Mitja ČOK
Dr. Tine STANOVNIK
Dr. Andreja CIRMAN
Mag. Valentina PREVOLNIK RUPEL
Dr. Mojmir MRAK

JAVNE FINANCE V SLOVENIJI

Ljubljana, 2007

Po uveljavitvi zakonodaje o davčnem postopku in davčni službi je postala vse nujnejša tudi zakonska ureditev ***davčnega svetovanja*** kot samostojne poklicne dejavnosti s podobnim položajem, kot je značilen za gospodarsko revizijo, odvetništvo ali notarstvo.

2. DOHODNINA (ZPOD)[1]

V Sloveniji je bila dohodnina uvedena l. 1991 in predstavlja prvi pomembnejši korak pri davčni reformi v obdobju prehoda. Znatne spremembe pri tem davku so nastopile leta 1994 in so zajele predvsem standardne in nestandardne olajšave, med drugim se je obseg nestandardnih olajšav zmanjšal od možnih 10% davčne osnove na vsega 3%, obseg olajšav za vzdrževanje družinskega člana pa se je povečal. Na osnovi odločbe Ustavnega sodišča RS iz leta 1996 so bile določbe zakona, ki se nanašajo na olajšave za vzdrževane družinske člane, v neskladju z ustavo. Preprosto povedano, te olajšave naj bi bile "preskromne". Obenem je Ustavno sodišče naložilo Državnemu zboru, da v roku enega leta to neskladnost odpravi. Enoletni rok se je krepko zavlekel in tako je bil novi zakon o dohodnini sprejet šele maja 2004, že kmalu po sprejetju je doživel prve spremembe, posebno velik »paket« sprememb zakona pa je bil sprejet v decembru 2005. Med njimi je najpomembnejša uvedba cedularne 20% obdavčitve za obresti, dividende in kapitalske dobičke. Od 1.1.2007 naprej pa velja popolnoma novi zakon o dohodnini, ki je poleg vrste ostalih sprememb prinesel tudi novo dohodninsko lestvico z zgolj tremi razredi in najvišjo mejno stopnjo 41%.

2.1 Vrste dohodkov, zavezanci in davčna osnova

Zakon o dohodnini določa naslednje vrste dohodkov:

1) dohodek iz zaposlitve,
2) dohodek iz dejavnosti,
3) dohodek iz osnovne kmetijske in osnovne gozdarske dejavnosti,
4) dohodek iz oddajanja premoženja v najem in iz prenosa premoženjske pravice,
5) dohodek iz kapitala,
6) drugi dohodki.

Med **dohodke iz zaposlitve** prištevamo poleg dohodka iz delovnega razmerja (plače, regres, bonitete, jubilejne nagrade, odpravnine, solidarnostne pomoči, nadomestila delodajalca, dohodki iz avtorskega dela, ustvarjenega iz delovnega razmerja, pokojnine, itd.) tudi dohodek iz drugega pogodbenega razmerja.

Za **dohodek iz dejavnosti** se šteje dohodek, dosežen z neodvisnim samostojnim opravljanjem dejavnosti. Med **dohodek iz osnovne kmetijske in osnovne gozdarske dejavnosti** pa se v osnovo všteva katastrski dohodek ter plačila iz naslova ukrepov kmetijske politike (subvencije) in iz naslova državne pomoči.

[1] Zbirka predpisov o dajatvah.

Za dohodek iz oddajanja premoženja v najem in iz prenosa premoženjske pravice se štejejo najemnine oz. dohodek za odstop pravice do uporabe oziroma izkoriščanja avtorskih pravic, izumov ipd.

Med **dohodek iz kapitala** štejejo obresti, dividende in kapitalski dobički. Ni vsak kapitalski dobiček obdavčen z dohodnino; obdavčitev je odvisna od oblike kapitala in od časa, ki je pretekel od pridobitve kapitala.

Drugi dohodki vključujejo nagrade, darila, priznavalnine, kadrovske in druge štipendije ipd.

Vsi našteti viri dohodnine se obdavčijo po poteku leta z dohodnino, med letom pa se plačuje akontacije dohodnine. Akontacija dohodnine torej omogoča sprotno pobiranje dohodnine in tako po poteku leta zneski doplačila oziroma vračila niso preveliki; s takšnim načinom se tudi zagotavlja redni mesečni priliv javnofinančnih prihodkov. V razdelku 2.4 so navedeni načini plačevanja akontacije dohodnine, prevladujoč način je v obliki davčnega odtegljaja (davka po odbitku). Pri nekaterih virih dohodnine je akontacijski davek proporcionalen: pri dohodkih iz zaposlitve je npr. za dohodek iz drugega pogodbenega razmerja stopnja 25%. Pri najpomembnejšem viru dohodnine, tj. plači, je akontacijski davek progresiven, in to v obliki stopničaste progresije, na mesečni ravni se uporablja lestvica, ki je enaka 1/12 letne; progresivna akontacijska lestvica velja tudi za pokojnine. Pri tem velja omeniti, da se za določene oblike dohodka iz dejavnosti (samostojni kulturniki, novinarji) akontacija dohodnine plačuje po proporcionalni stopnji 25%. Za zavezance za davek od dohodka iz dejavnosti, za katere se davčna osnova ugotavlja na podlagi dejanskih prihodkov in odhodkov, pa velja, da si sami izračunajo akontacijo dohodnine od dohodka iz dejavnosti v davčnem obračunu. Za akontacijo davka od dohodka iz osnovne kmetijske in osnovne gozdarske dejavnosti je stopnja 10%, če je skupni znesek katastrskega dohodka na zavezanca enak ali večji od 10% povprečne plače v Sloveniji za preteklo leto. Kmetje, ki tega kriterija ne dosegajo, ne plačujejo akontacijskega davka.

Zavezanci za dohodnino so fizične osebe – rezidenti in nerezidenti. Razlikovanje je pomembno zaradi davčne obveznosti: rezidenti so obdavčeni po načelu svetovnega dohodka, nerezidenti pa samo po načelu vira dohodka, tj. obdavčen je samo dohodek, ki ga ustvarijo v Sloveniji. Za rezidenta se šteje oseba, ki ima uradno prijavljeno stalno prebivališče v Sloveniji. Rezidenti so tudi osebe, ki imajo (a) svoje običajno bivališče ali središče svojih osebnih in ekonomskih interesov v Sloveniji ali (b) so v Sloveniji prisotni skupno več kot 183 dni v letu.

Do neto letne **davčne osnove dohodnine** pridemo tako, da od vsote davčnih osnov posameznih vrst dohodkov (tj. "dohodkov, ki so predmet davčne obravnave") odštejemo prispevke za socialno varnost, priznane stroške ter olajšave. Zakon natančno določa, kaj so

priznani stroški; pri nekaterih dohodkih, ki so predmet davčne obravnave – tu mislimo predvsem na dohodke iz dejavnosti, pa zakon sproti navaja tudi olajšave.

1. ***Dohodek iz zaposlitve:***
   (a) *Letna bruto plača manj obvezni prispevki za socialno varnost.* Med "plače" se za obdavčitev štejejo tudi: a) letni zneski *nadomestil plač*, b) letni zneski *regresa*, c) tisti del *povračil stroškov v zvezi z delom* (dnevnic, stroškov prevoza na delo, prehrana med delom, povračilo stroškov za prenočišče, »kilometrina«), ki presega višino, določeno s strani Vlade RS, d) jubilejne nagrade, ki presegajo višino, določeno s strani Vlade RS, in odpravnine, ki presegajo znesek, določen v zakonu o dohodnini, e) letni zneski *bonitet*, ki jih zagotavlja delodajalec (tu gre predvsem za uporabo službenega avtomobila v privatne namene, zagotavljanje nastanitve, zagotavljanje ugodnega posojila ipd).
   Vidimo torej, da se priznavajo le zelo določene oblike stroškov, ki so nastali ob pridobivanju dohodka, pa še ti stroški se priznavajo le do dovoljene višine.
   (b) *Dohodek iz drugega pogodbenega razmerja.* Priznavajo se normirani stroški v višini 10% bruto prejemka. Poleg normiranih stroškov je mogoče uveljavljati tudi dejanske stroške prevoza in nočitev v zvezi z opravljanjem dela, in sicer do višin, ki jih določi vlada.
   (c) Letni znesek *pokojnine.*

2. ***Dohodek iz dejavnosti:***
   Dohodek od dejavnosti, tj. prihodki manj odhodki. Stroški, ki so nastali ob pridobivanju dohodka, se priznavajo skoraj v celoti; obstaja le nekaj manjših omejitev za odhodke (plačani davki, samoprispevki). Med nepriznane stroške prištevamo tudi stroške nakupa delnic ali posojanja; zato se tudi pri prihodkih ne vštevajo dividende in obresti. Nekateri stroški, npr. stroški za uporabo avtomobila, ne smejo preseči zneska, določenega s predpisi Vlade RS. Med odhodke se prav tako vštevajo tudi obvezni prispevki za socialno varnost, ki se nanašajo na zavezanca, ne pa tudi "lastna plača" oziroma dohodek zavezanca - zato se z vidika ekonomske terminologije ne obdavčuje dobiček zavezanca, temveč njegov dohodek. Zavezancu se priznajo različne olajšave (kot odbitek pred obdavčljivo osnovo), predvsem za vlaganja lastnih sredstev v raziskave in razvoj, zaposlovanje določenih kategorij delavcev itd. V bistvu gre za podobne olajšave, kot so jih deležne pravne osebe, zavezanci za davek od dohodka pravnih oseb. V primeru izgube ima zavezanec možnost izgubo pokrivati z zmanjšanjem davčne osnove od dohodka iz dejavnosti v naslednjih letih. Zavezanci, ki ne zaposlujejo delavcev in imajo letne prihodke nižje od 42.000 EUR, lahko zahtevajo poenostavljeno davčno obravnavo, in sicer tako, da se odhodki določijo normativno, v višini 25% ustvarjenih prihodkov. Pri tej skupini se torej davčna osnova določi tako, da se od prihodkov odštejejo prispevki za socialno varnosti in normirani odhodki.

3. ***Dohodki iz osnovne kmetijske in osnovne gozdarske dejavnosti:***

Dohodek vključuje (a) katastrski dohodek od kmetijskih in gozdnih zemljišč in potencialne tržne dohodke od pridelave v panjih ter (b) plačila iz naslova ukrepov kmetijske politike in druga plačila iz naslovov državnih pomoči. Kot je znano, je katastrski dohodek normativni koncept, ker se ne določa na osnovi dejanskega dohodka. Pri tem dohodku se zato ne priznavajo stroški, ki so nastali ob pridobivanju dohodka, toda zelo velikodušno se priznavajo različne oprostitve, ki so vezane predvsem na to, da se katastrski dohodek »nizkokvalitetnega« zemljišča ne všteva v davčno osnovo.

4. ***Dohodek iz oddajanja premoženja v najem in iz prenosa premoženjske pravice:***

(a) Dohodki od *oddajanja premoženja v najem.* Davčna osnova je dohodek, zmanjšan za normirane stroške v višini 40% najemnine; namesto tega pa zavezanec lahko uveljavlja dejanske stroške vzdrževanja premoženja.

(b) *Dohodek iz prenosa premoženjske pravice*, tj. odstop pravice do uporabe oziroma izkoriščanja avtorskih pravic, izumov ipd. Davčna osnova je dohodek, zmanjšan za normirane stroške, v višini 10% dohodka.

5. ***Dohodek iz kapitala:***

(a) *Obresti.* Obresti iz določenih naslovov niso obdavčene, tj. za njih velja oprostitev. Še pomembneje pa je to, da se obresti od denarnih depozitov ne obdavčijo, če ne presegajo zneska 1.000 EUR (to je neobdavčen del obresti).

(b) *Dividende.*

(c) *Kapitalski dobički.* Obdavčitev kapitalskih dobičkov je v sistemih dohodnine »najtrši« oreh. Tu je potrebno določiti, katere oblike kapitala so sploh relevantne, ko določamo kapitalske dobičke, kakšne so časovne omejitve, kako natančno merimo kapitalske dobičke in v katerih primerih se le-ti obdavčijo. Tako se kapitalski dobički obdavčijo ob odsvojitvi, ki vključuje poleg prodaje tudi dajanje v dar in zamenjavo kapitala. Z dohodnino se obdavčujejo kapitalski dobički, ki izvirajo iz nepremičnin, vrednostnih papirjev in drugih deležev v gospodarskih družbah ter investicijskih kuponov. Davčna osnova je razlika med vrednostjo kapitala ob odsvojitvi in vrednostjo kapitala ob pridobitvi, pri čemer zakon tudi zelo natančno določa, kaj je razumljeno pod tema dvema pojmoma. Začetna cedularna davčna stopnja 20% se znižuje za ¼ za vsakih dopolnjenih pet let lastništva kapitala. Edina izjema so stanovanja, v katerih ima zavezanec stalno prebivališče – v tem primeru se kapitalski dobiček obdavči samo, če je realiziran v manj kot treh letih od pridobitve stanovanja.

6. ***Drugi dohodki:***

Drugi dohodki vključujejo heterogeno skupino »ostankov«, tj. dohodkov, ki jih ne moremo razvrstiti v zgoraj navedene skupine dohodkov, prav tako pa jih zakon izrecno ne razvršča v skupino dohodkov, ki so oproščeni plačila dohodnine, niti ne v skupino dohodkov, ki se ne vštevajo v davčno osnovo. Drugi dohodki vključujejo: nagrade, darila, dobitke v nagradnih igrah, priznavalnine, kadrovske in druge štipendije, izplačila odkupne vrednosti pokojninskega zavarovanja, itd. Obstajajo številne olajšave in oprostitve pri tovrstnih dohodkih. Tako se npr. v davčno osnovo ne všteva kadrovska in druga štipendija do višine minimalne plače; izplačila odkupne vrednosti v primeru rednega prenehanja prostovoljnega dodatnega pokojninskega zavarovanja pa niso obdavčena, če se sredstva prenesejo v zavarovalnico, ki potem izplačuje zavarovancu mesečne doživljenjske rente.

2.2 Olajšave

Večina olajšav je določene v absolutnem znesku in se usklajujejo enkrat letno z rastjo cen življenjskih stroškov.

a) ***Splošna olajšava:*** vsakemu zavezancu se osnova zmanjša za 2.800 EUR.

b) ***Osebne olajšave*** znašajo: (1) invalidom s 100% telesno okvaro in pravico do tuje nege in pomoči se osnova zniža za 14.971 EUR; (2) zavezancem po dopolnjenem 65. letu starosti se osnova zniža za 1.205 EUR. (3) prejemnikom pokojnin iz obveznega pokojninskega in invalidskega zavarovanja, prejemnikom nadomestil iz naslova obveznega invalidskega zavarovanja in prejemnikom priznavalnin se prizna davčna olajšava v obliki znižanja davčne obveznosti v višini 13,5% odmerjenega prejemka;

c) ***Posebne osebne olajšave*** so namenjene delovno aktivnim zavezancem, in vključujejo (1) samozaposlene v kulturi in (2) samostojne novinarje, pa tudi (3) dijake in študente za dohodek, ustvarjen prek študentskih servisov. Samozaposlenim v kulturi in samostojnim novinarjem se, pod pogojem, da so vpisani v ustrezen »razvid«, prizna zmanjšanje davčne osnove v višini 15% prihodkov letno, do zneska 25.000 EUR prihodkov letno. Če torej zavezanec ustvari 40.000 letnih prihodkov, se mu prizna znižanje davčne osnove samo v višini 3.750 EUR (15% od 25.000 EUR). Vse te olajšave veljajo samo za zavezance, ki nimajo sklenjenega delovnega razmerja. Študentom se prizna davčna olajšava v obliki znižanja davčne osnove za 2.800 EUR letno.

d) ***Posebne olajšave*** so namenjene zavezancem, ki vzdržujejo družinske člane: za prvega otroka in vsakega drugega vzdrževanega družinskega člana se prizna davčna olajšava v višini 2.066 EUR letno. Za vsakega nadaljnjega otroka se olajšava poveča. Za otroka z motnjami pa se prizna davčna olajšava v višini 7.486 EUR letno. Za otroke, ki se šolajo na srednji, višji ali visoki stopnji ter niso v delovnem razmerju, velja olajšava največ do

26. leta starosti (pod pogojem, da »pavzirajo« največ eno leto). Olajšava po 26. letu se priznava le tistim študentom, ki so se vpisali na študij do 26. leta starosti in največ za dobo 6 let od vpisa.

e) ***Olajšava za prostovoljno dodatno pokojninsko zavarovanje***. Zavezanec, ki zase plačuje premijo prostovoljnega dodatnega pokojninskega zavarovanja (pri čemer mora biti pokojninski načrt odobren in vpisan v poseben register), se prizna olajšava v višini 24 % obveznih prispevkov za pokojninsko in invalidsko zavarovanje, vendar največ do zneska 2.390 EUR letno. Če delodajalec in delojemalec oba plačujeta premije za prostovoljno dodatno pokojninsko zavarovanje za zavezanca, ima pri določitvi višine olajšave za plačane premije delodajalec »prednost«. To pomeni, da če delodajalec vplača 2.000 EUR letno za zavezanca, sam zavezanec pa 1.000 EUR letno, bo zavezanec lahko uveljavljal največ 390 EUR davčne olajšave (2.390 EUR - 2.000 EUR).

2.3 Lestvica, razredi in progresija

Lestvica: stopničasta progresija, ki se valorizira vsako leto glede na rast cen življenjskih potrebščin.

Število in širina razredov: trije razredi: 0 - 6.800 / 6.800 - 13.600 / nad 13.600 EUR.

Stopnje: 16%, 27% in 41%.

Cedularna obdavčitev: obresti, dividende in kapitalski dobički so obdavčeni z 20% stopnjo.

Popolnost zajetja dohodkov: v primerjavi z dohodninami v razvitih deželah se v Sloveniji v davčno osnovo vključujejo skoraj vse pomembnejše oblike dohodkov fizičnih oseb. Pomembnejše izjeme so le a) "pripisani dohodki" zaradi prebivanja v lastnem stanovanju, b) vse oblike socialnih pomoči (ponekod v tujini obdavčujejo zlasti otroške dodatke). Zato pa so pri nas vključeni v obdavčitev tudi t.i. *"stimulacije in bonitete iz delovnega razmerja"* (*fringe benefits*) ter *boleznine in druga nadomestila* (npr. nezaposlenim), kar je v kar precej razvitih deželah neobdavčeno ali vključeno med olajšave. Prav tako ne gre prezreti, da so določeni dohodkovni viri deležni kar občutnih olajšav (npr. dohodki študentov, kulturnikov ipd.)

Z vidika resnične prakse je popolnost zajetja dohodkov v obdavčitev precej problematična: zelo slab nadzor omogoča obsežno utajevanje dohodkov iz "sive ekonomije". Obdavčitev torej zajema predvsem tista izplačila prejemkov, ki so evidentirana pri pravnih osebah.

2.4 Načini plačevanja dohodnine

Obstajajo trije načini medletnega plačevanja davkov. Ti so:

(a) ***Davčni odtegljaj (davek po odbitku).***

Takšna oblika obdavčitve se uporablja predvsem pri tistih dohodkih oziroma izplačilih, ki so "vidni" oziroma pri katerih je možnost nadzora enostavna. Tukaj gre predvsem za dohodke fizičnih oseb, ki jih izplačujejo pravne osebe in osebe, ki opravljajo dejavnost: plače, pokojnine, dohodek iz drugega pogodbenega razmerja, dohodki iz oddajanja premoženja v najem, nekatere oblike dohodka iz dejavnosti (predvsem tistih, pri katerih se davčna osnova ugotavlja na osnovi normiranih odhodkov). Pri plačah in pokojninah je davčni odtegljaj dokaj natančen, upošteva dejansko dohodninsko lestvico, tako da naj bi bila po preteku leta doplačila ali povračila dohodnine čim manjša. Pri plačilih na osnovi drugega pogodbenega razmerja in pri dohodkih od prenosa premoženjske pravice ter dohodkih iz oddajanja premoženja v najem so ti akontacijski davki proporcionalni in znašajo 25% osnove. Davčni odtegljaj se prav tako uporablja za določene skupine zavezancev za davek od dohodka iz dejavnosti, kot npr. pri samostojnih kulturnih delavcih, samostojnih novinarjih ipd. Tudi pri teh skupinah zavezancev je stopnja davčnega odtegljaja 25%, osnova pa je prav tako izplačilo, zmanjšano za priznane stroške. Naj omenimo, da so za določene skupine (npr. za zavezance za davek od dohodka iz dejavnosti, ki ugotavljajo davčno osnovo na podlagi normiranih stroškov, možna tudi nižja akontacijska stopnja od 25%.

Pri dohodkih, ki so na letni ravni obdavčeni cedularno, medletno plačilo 20% davka predstavlja že dokončno letno davčno obveznost.

(b) ***Davek, plačan na osnovi odločbe o odmeri davka.***

Takšna oblika obdavčitve je značilna predvsem za dohodke, ki jih izplačujejo fizične osebe, in za dohodke, prejete iz tujine – izplačevalci teh dohodkov niso plačniki davka, zato za davčnega zavezanca niso dolžni obračunati in plačati davek. Prejemniki teh dohodkov morajo zato vložiti davčno napoved praviloma v 15 dneh, a ne vedno v tem roku, po prejemu dohodka pri davčnem organu, davčni organ pa izda odločbo o odmeri davka. To npr. velja za dohodek iz oddajanja premoženja v najem in za dobiček pri odsvojitvi nepremičnine. Zavezanci za davek od dohodka iz dejavnosti, ki uveljavljajo normirane stroške, morajo vložiti davčno napoved do 15. dne v mesecu za preteklo trimesečje. Pri določenih oblikah dohodka, npr. pri najemninah, kjer gre za celoletni kontinuiran dohodek, in dohodku iz kapitala, ki vključuje obresti, dividende in dobiček iz kapitala, se davčna napoved vloži v določenih rokih prvega trimesečja tekočega leta za preteklo leto. Ob tem lahko omenimo, da napovedi za davek od dohodka iz kmetijstva in gozdarstva ni potrebno vlagati, ker ti dohodki temeljijo na pavšalu, to je na katastrskem dohodku. Davčni organ na tej podlagi izda odločbo o odmeri akontacije dohodnine od katastrskega dohodka, ki se plača v treh obrokih.

(c) ***Samoobdavčitev.***

To je novejša oblika obdavčevanja, ki je stopila v veljavo s 1.1.1998 in velja za zavezance za davek od dohodka iz dejavnosti, ki ugotavljajo davčno osnovo na osnovi dejanskih prihodkov in odhodkov. Zavezanec si mora sam izračunavati akontacijo davka iz dejavnosti v davčnem obračunu, saj mu akontacija davka iz dejavnosti ni predpisana z odločbo davčnega organa.

2.5 Vložitev letnih napovedi

V zvezi s plačevanjem dohodnine je treba omeniti tudi davčni obračun oz. letne napovedi za posamezne vire dohodnine. Tako zavezanec za davek od dohodka iz dejavnosti predloži davčni obračun do 31. marca tekočega leta za preteklo leto. Napoved za odmero davka od obresti na denarne depozite ter dobička iz kapitala, če je dosežen s prodajo vrednostnih papirjev ter drugih deležev v kapitalu ter investicijskih kuponov, mora davčni zavezanec vložiti do 28. februarja tekočega leta za preteklo leto. Zavezanec, ki oddaja premoženje v najem za več mesecev v letu, mora oddati napoved za odmero davka do 15. januarja tekočega leta za preteklo leto.

2.6 Odmera dohodnine

V letu 2007 je zakon določil tudi drugačen način odmere dohodnine. Davčni organ sestavi za davčnega zavezanca rezidenta do 31. maja tekočega leta za preteklo leto informativni izračun dohodnine, ki se šteje za njegovo davčno napoved, če davčni zavezanec ne ugovarja v 15 dneh od dneva vročitve informativnega izračuna dohodnine. Po poteku roka za ugovor, in če zavezanec rezident ni ugovarjal, postane informativni izračun dohodnine odločba o odmeri dohodnine.

Če davčni zavezanec ugovarja zoper informativni izračun, v roku za ugovor vloži dopolnjen informativni izračun, ki se šteje za njegovo napoved za odmero dohodnine. V teh primerih davčni organ preveri navedbe davčnega zavezanca ter na podlagi svojih podatkov in podatkov davčnega zavezanca izda odločbo.

3. DAVEK NA DODANO VREDNOST (ZPOD)[2], TROŠARINE (ZPOD) IN DAVEK NA MOTORNA VOZILA (ZPOD)

3.1 Davek na dodano vrednost

Republika Slovenija (RS) je uvedla davek na dodano vrednost (DDV) s 1. julijem 1999, ko je z njim nadomestila davek od prometa proizvodov in storitev. DDV je v celoti prihodek proračuna RS.

S 1.5.2004 je RS postala članica Evropske unije (EU) in s tem se je v nekaterih elementih spremenil tudi sistem DDV, saj je del nalog, ki jih je do vstopa opravljala carina, odpadel na davčne zavezance, drugačna pa so tudi pravila pri obračunavanju DDV v prometu z drugimi članicami EU. Za promet z drugimi članicami EU na enotnem trgu tako uporabljamo izraz *znotrajskupnostni* ali *intrakomunitarni promet* (intrakomunitarna dobava v drugo državo EU in intrakomunitarna pridobitev iz druge države EU).

Pri prometu med davčnimi zavezanci velja načelo ***države uvoznice*** (blago je obdavčeno v državi kamor je uvoženo). Ko slovenski zavezanec kupi blago od zavezanca iz druge države EU, blago zapusti državo izvoznico brez da bi bilo obremenjeno z njenim DDV (obdavčeno je s stopnjo 0% DDV). Slovenski uvoznik potem sam obračuna DDV (po slovenski stopnji), pri čemer si tako obračunani DDV odbije kot vstopni DDV (***metoda obračuna po sedežu prejemnika***). Pogoj za aplikacijo načelo države uvoznice pa je, da ima kupec veljavo identifikacijsko številko za DDV ter ustrezno prevozno listino, s katero dokaže, da je blago zapustilo državo porekla ter prispelo v RS. Načelo države uvoznice velja tudi za promet med davčnimi zavezanci in pravnimi osebami, ki niso davčni zavezanci (če letna vrednost prometa presega 10.000 EUR), velja pa tudi za dobave končnim potrošnikom, kmetom pavšalistom ter pravnim osebam, ki niso davčni zavezanci pri pošiljčnih poslih (prodaji na daljavo), če letna vrednost prometa presega 35.000 EUR, ter za promet novih vozil in trošarinskega blaga.
V vseh ostali primerih, ko je kupec končni potrošnik pa velja ***načelo države izvoznice***, blago se obdavči po predpisih države, kjer se blago ob nakupu nahaja. Če slovensko podjetje proda blago fizični osebi iz druge države EU, bo kupec iz tujine plačal celotni DDV v RS.

[2] Zbirka predpisov o dajatvah.

3.1.1. Dobava

DDV se obračunav[a]h storitev, ki jih davčni zavezanec opravi v okviru op[ra]ter od uvoza blaga/storitev. Uvoz blaga/storitev pom[e]h držav.

Dobava blaga p[o] opredmetimi stvarmi. Za dobavo blaga se šteje pr[i] opravljen za plačilo na podlagi odločbe državnega [or]godbe o najemu (lizing), izročitev novozgrajenih obj[e]tvi davčnega zavezanca s strani druge osebe, do[b]rgije za ogrevanje ali hlajenje, zamenjava blaga, [n]eposlovne namene in brezplačna odtujitev oz. odtu[je]ja poslovnih vzorcev v razumnih količinah ipd.), jem

Opravljanje storitev je dobava blaga. Tako se za opravljanje storitev šteje tudi [i]nje storitev na podlagi odločbe državnega organa, [z]a neposlovne namene, zamenjava storitev itd.

Z DDV pa ni obdav[e]zičnimi osebami.

3.1.2 Plačnik[o]v dogodek in nastanek obvez[n]osti

Plačnik DDV

Plačnik davka je ***d***[o]promet blaga/storitev, od katerega se obračunava DD[V] RS nima sedeža, pa je plačnik njegov davčni zas[t]zavezanec ne imenuje svojega zastopnika. Plačnik[a]V, ki ga po zakonu ne bi smela izkazati, pri uvozu [s]carinskimi predpisi oz. prejemnik uvoženega blaga.

Kraj obdavčitve

DDV se obračunav[a]jen promet blaga/storitev. Če ni drugače določeno, [g]ometa. Za kraj dobave blaga se šteje kraj, kjer se [p]ošiljanje ali prevoz, kraj, kjer se blago instalira oz. [k]enega prometa blaga šteje kraj, kjer je blago vneš[e] praviloma šteje kraj, kjer ima zavezanec, ki opravlja storitve, svoj sedež ali poslovno enoto, kraj, kjer leži nepremičnina, če se opravljajo storitve v zvezi z nepremičnino, kraj, kjer ima prejemnik storitve iz RS sedež, če storitev opravi zavezanec, ki v RS nima sedeža, itd.

V primeru pošiljalčnih poslov oz. prodaje na daljavo, ko davčni zavezanec iz države EU brez neposrednega stika s kupci iz RS po pošti, železnici ipd., dobavi blago končnim potrošnikom, kmetom pavšalistom ter pravnim osebam, ki niso davčni zavezanci, pa je takšno blago obdavčeno zgolj v državi izvoznici, ob pogoju, da skupna vrednost takšnega blaga v tekočem koledarskem letu ne presega 35.000 EUR niti ni tega zneska presegla v preteklem letu. Če je mejni znesek 35.000 EUR presežen, se mora izvoznik registrirati v RS (izjema od tega pravila so nova prevozna sredstva in trošarinski izdelki, v teh primerih je kraj nastanka davčne obveznosti vedno RS).

Obdavčljiv dogodek in nastanek obveznosti obračuna DDV

Obdavčljiv dogodek in s tem obveznost obračuna DDV nastane, ko je blago dobavljeno ali ko so storitve opravljene. Če je račun izdan pred opravljeno dobavo blaga oz. opravljeno storitvijo nastane obveznost obračuna DDV, ko je izdan račun *(načelo fakturirane realizacije oz. načelo nastanka dogodka).* Če račun ni bil izdan pa najkasneje zadnji dan davčnega obdobja, v katerem je nastal obdavčljiv dogodek.

Za dobave blaga in opravljanje storitev, pri katerih se izdajajo zaporedni računi ali se opravljajo zaporedna plačila, se šteje da so opravljene v trenutku, ko poteče obdobju, na katero se taki obračuni ali plačila nanašajo, vendar to obdobje ne sme biti daljše od enega leta. V primeru vnaprejšnjega plačevanja nastane obveznost obračuna DDV na dan sprejema plačila in od prejetega zneska plačila.

Davčni zavezanci, katerih letni promet je manjši od 208.000 EUR pa se lahko odločijo za obračun DDV na podlagi *plačane realizacije (načelo denarnega toka).*

Pri ***uvozu iz tretjih držav*** nastane obveznost obračuna DDV ob uvozu, takrat ko nastane tudi obveznost za obračun uvoznih dajatev.

3.1.3. Davčna osnova, stopnje, oprostitve

Davčna osnova je vse, kar predstavlja plačilo (v denarju, stvareh ali storitvah), ki je bilo zaračunano kupcu, povečano za trošarine in druge davke, takse in dajatve (npr. carine) ter posredne stroške (provizije, stroški embaliranja, prevoza in zavarovanja), ki jih zaračuna dobavitelj. V davčno osnovo se ne štejejo znižanja cen in popusti, odobreni v plačilu.
Pri ***uvozu*** je osnova vrednost blaga, določena v skladu s carinskimi predpisi.

V slovenskem sistemu sta dve stopnji DDV:

a) Standardna v višini 20%, s katero je obdavčeno vse blaga in storitve, ki niso oproščeni ali obdavčeni z nižjo stopnjo.

b) Nižja stopnja v višini 8,5%, ki se obračunava od:

1. Hrane (vključno s pijačo, razen alkoholnih pijač), za ljudi in živali, živih živali, semen, rastlin in primesi ter dodatkov, ki se običajno uporabljajo kot dodatki k hrani, ter priprava jedi (gostinska dejavnost).
2. Dobave vode.
3. Zdravil, vključno z izdelki, ki se uporabljajo za kontracepcijo ter sanitarno zaščito.
4. Medicinske, ortopedske in rehabilitacijske opreme oz. pripomočkov, vključno z vzdrževanjem te opreme in pripomočkov.
5. Prevoza potnikov in njihove prtljage.
6. Knjig, časopisov in periodičnih publikacij.
7. Vstopnin za razstave, gledališča, za ogled naravnih znamenitosti, kinematografske in glasbene prireditve, sejme, cirkuse, živalske vrtove ipd. ter vstopnic za športne prireditve.
8. Avtorskih pravic književnikov in skladateljev ter storitev izvajalskih umetnikov.
9. Uvoza umetniških predmetov ter zbirk in starin.
10. Stanovanj, stanovanjskih in drugih objektov, namenjenih za trajno bivanje, ter delov teh objektov, če so del socialne politike, vključno z gradnjo, obnovo in popravili le-teh.
11. Živali za pitanje, semen sadik, gnojil, fitofarmacevtskih sredstva ter storitev, ki so namenjene izključno za uporabo v kmetijstvu, gozdarstvu in ribištvu.
12. Nastanitev v hotelih, penzionih, domovih, kampih in podobnih nastanitvenih zmogljivostih ter oddajanja prostorov za kampiranje.
13. Uporabe športnih objektov.
14. Storitev pokopa in upepelitve, skupaj s prometom blaga, povezanega s temi storitvami.
15. Storitev javne higiene.

Oprostitev velja za **(a) dejavnosti, ki so v javnem interesu.** Sem spadajo:

1. Zdravstvene storitve in oskrba, vključno z oskrbo s človeškimi organi, krvjo in materinim mlekom.
2. Storitve, ki jih za svoje člane opravljajo neodvisne skupine oseb, katerih dejavnosti so oproščene plačila DDV ali so neobdavčljive, če te skupine od svojih članov zahtevajo le plačilo njihovega deleža skupnih stroškov in ni verjetno, da bi takšna oprostitev vodila k izkrivljanju konkurence.
3. Socialno-varstvene storitve in dobava blaga, neposredno povezanega s temi storitvami.
4. Storitve otroškega varstva, vzgoje in izobraževanja (na vseh ravneh), vključno z dobavami blaga in storitev, neposredno povezanega s temi storitvami.

5. Zagotavljanje osebja, s strani verskih skupnosti ali filozofskih združen za opravljanje dejavnosti iz 1., 3. in 4. točke.
6. Opravljanje storitev in dobava blaga, neposredno povezanega s temi storitvami, ki jih opravljajo nepridobitne politične, sindikalne, verske, rodoljubne, dobrodelne in podobne organizacije dobavljajo svojim članom kot povračilo za članarino in ni verjetno, da bi takšna oprostitev vodila k izkrivljanju konkurence.
7. Storitve, povezane s športom in športno vzgojo, ki jih posameznikom nudijo neprofitne organizacije.
8. Kulturne storitve in z njimi neposredno povezano blago, ki jih opravljajo javni zavodi in druge s strani države priznane kulturne organizacije.
9. Opravljanje storitev in dobava, ki ga opravljajo organizacije, katerih dejavnosti so oproščene DDV po 1., 3., 4., 6., 7. in 8. točki, če se ta promet opravlja v povezavi z zbiranjem denarnih sredstev, ki ga te organizacije opravljajo izključno v lastno korist in ni verjetno, da bi takšna oprostitev vodila k izkrivljanju konkurence.
10. Reševalni prevozi oseb v za ta namen posebej prirejenih vozilih.
11. Prispevek za programe RTV Slovenija.

Oprostitev velja tudi za **(b) druge dejavnosti**, kamor spadajo:

1. Zavarovalni in pozavarovalni posli, vključno s povezanimi storitvami, ki jih opravljajo zavarovalniški posredniki in zastopniki.
2. Najem oz. zakup nepremičnin (vključno z lizingom) razen: nastanitev v hotelih, kampih ipd., oddajanja parkirnih prostorov in garaž, oddajanja trajno instalirane opreme in strojev ter najema sefov.
3. Finančne storitve (odobravanje in upravljanje kreditov, storitve v zvezi z upravljanjem depozitov, hranilnih vlog, bančnih računov, opravljanjem plačilnega prometa, promet z delnicami in drugimi vrednostnimi papirji, upravljanje z investicijskimi skladi, itd.).
4. Kolki in druge podobne znamke.
5. Igre na srečo.
6. Dobava "starih" objektov (nepremičnin) in zemljišč, na katerih so ti objekti postavljeni.
7. Dobava zemljišč, razen stavbnih zemljišč.
8. Dobava zlata Banki Slovenije.

Plačila davka je oproščeno tudi blago, ki je v tranzitu čez RS, blago, ki je predmet odloženih carinskih postopkov, blago, ki je oproščeno carine in je namenjeno diplomatskim predstavništvom ter tujim mednarodnim organizacijam, če tako določajo mednarodne pogodbe, itd.

Čeprav zakon govori o oprostitvi izvoza, je **izvoz blaga/storitev** dejansko obdavčen s **stopnjo 0%**, saj imajo izvozniki pravico do odbitka vstopnega (nakupnega) DDV od vseh nabav, ki so potrebne za realizacijo izvoza.

3.1.4 Izdaja računov, davčno obdobje, obračunavanje in plačevanje davka, vključitev v sistem DDV

Davčni zavezanec mora za vsako dobavo blaga/storitev izdati ***račun*** (lahko tudi v nematerializirani obliki). Če je le-ta namenjen davčnim zavezancem, mora vsebovati datum izdaje in zaporedno številko računa, identifikacijsko številko izdajatelja računa in kupca, ime in naslov dobavitelja in kupca, opis blaga/storitev, datum odpošiljanja blaga/opravljanja storitev, ceno blaga/storitve brez DDV, ter znesek DDV po različnih stopnjah.

Davčno obdobje, obračunavanje in plačevanje davčne obveznosti

Davčno obdobje, v katerem mora zavezanec obračunavati DDV, je odvisno od obdavčljivega prometa v preteklem koledarskem letu. Če je promet presegal 210.000 EUR, je davčno obdobje en mesec, sicer pa je davčno obdobje trimesečje.

Izvozniki, ki dobavljajo blago pretežno v druge države EU, imajo ne glede na obseg prometa možnost, da DDV obračunavajo mesečno.

Davčni zavezanec mora obračun davčne obveznosti, t.j. razlike med izstopnim in vstopnim DDV davčnega obdobja, predložiti do zadnjega delovnega dne naslednjega meseca po poteku davčnega obdobja ne glede na to, ali je v navedenem obdobju dolžan plačati DDV.

Zavezancu, ki ima do države ***terjatev***, t.j. znesek vstopnega DDV je večji od zneska izstopnega DDV, se razlika všteje v plačilo v naslednjih davčnih obdobjih ali pa se na zahtevo zavezanca vrne v 30. - 60. dneh.

Davčni zavezanec, identificiran za namene DDV mora davčnemu organu trimesečno poročati o vseh dobavah blaga, ki jih opravi osebam identificiranim za namene DDV v drugih državah EU (za potrebe preverjanj znotraj sistema VIES – VAT Information Exchange System).

Če zavezanec opravlja obdavčljivo in oproščeno dejavnost (ter zanju ne vodi ločenega knjigovodstva), določi znesek vstopnega DDV, ki se nanaša na obdavčljivo dejavnost z ***odbitnim deležem***; v številu je znesek obdavčljivega letnega prometa, zmanjšan za DDV, ki se nanaša na ta promet, v imenovalcu pa je celotni znesek letnega prometa, zmanjšan za DDV, ki se nanaša na ta promet. Odbitni delež za tekoče leto se začasno določi na podlagi podatkov o prometu preteklega leta.

Kupec, ki je fizična oseba in v EU nima prebivališča, ima v roku treh mesecev pravico do vračila DDV od blaga, nabavljenega v RS, ki ga iznese iz EU (razen za naftne derivate, alkoholne pijače in tobačne izdelke), če znaša vrednost nakupa nad 50 EUR.

3.1.5 Vključitev v sistem DDV, posebni postopki obdavčenja

Prag za ***obvezno vključitev v sistem DDV*** je pri vrednosti prometa blaga/storitev, ki presega 25.000 EUR v zadnjih dvanajstih mesecih (7.500 EUR za kmetije, katerih katastrski dohodek presega ta znesek), vsi ostali zavezanci se lahko vključijo v sistem prostovoljno, vendar najmanj za obdobje petih let.

Posebni postopki obdavčenja

a) ***Kmetje***, ki ne dosegajo praga za obvezno vključitev v sistem DDV (7.500 EUR katastrskega dohodka) ter prodajajo izdelke kupcem, ki so v sistemu, so upravičeni do t.i. *pavšalnega nadomestila vstopnega DDV* v višini 4% prodajne vrednosti. Na prodajno vrednost svojih pridelkov zaračunajo dodatne 4% *pavšalnega nadomestila,* s katerimi si "povrnejo" plačan vstopni DDV, to *pavšalno nadomestilo* pa se pri kupcih kmetijskih izdelkov upošteva kot njihov vstopni DDV. Za vključitev v takšno shemo obdavčenja potrebujejo dovoljenje davčnega organa.

b) **Potovalna agencija**, ki deluje v svojem imenu, nekatere storitve povezane s potovanjem pa zaupa drugim izvajalcem, obračuna DDV po »direktni odštevalni metodi«. DDV je obračunan od razlike med zneskom, ki ga plača potnik v katerega ni vključen DDV in dejanskimi stroški potovalne agencije za storitve, ki jih opravijo drugi izvajalci, če je neposreden uporabnik teh storitev potnik. Potovalna agencija tako v nobeni državi EU nima pravica do odbitka vstopnega DDV, niti do vračila DDV, ki ji ga zaračuanjo drugi izvajalci od storitev, ki so jih nudili neposredno potniku. Če pa so storitve, ki jih izvedejo drugi izvajalci opravljene zunaj EU, se storitev potovalne agencije šteje za oproščeno dejavnost.

c) ***Davčni zavezanec - preprodajalec***, ki pridobiva rabljeno blago, umetniške predmete, zbirke in starine z namenom nadaljnje preprodaje, obračunava DDV od ustvarjene razlike med prodajno ceno in nabavno ceno, zmanjšane za znesek DDV, ki se nanaša na razliko v ceni. Za vstop v shemo se lahko preprodajalec odloči prostovoljno, toda najmanj za obdobje dveh let.

3.2. Trošarine

Trošarina se plačuje od trošarinskih izdelkov (alkohola in alkoholnih pijač, tobačnih izdelkov ter energentov in električne energije), ki se na območju RS sprostijo v uporabo.

Trošarine so prihodek proračuna RS.

3.2.1. Osnovni pojmi, obračun trošarin

Trošarinski zavezanec je proizvajalec trošarinskih izdelkov, pooblaščeni prejemnik trošarinskih izdelkov iz druge države EU, uvoznik trošarinskih izdelkov, trgovec na debelo s trošarinskimi izdelki oz. oseba, na katero se prenese trošarinska obveznost v skladu z zakonom. Imetnik trošarinskega dovoljenja je trošarinski zavezanec, ki pridobi ***trošarinsko dovoljenje*** carinskega organa, da lahko v okviru opravljanja svoje dejavnosti v trošarinskem skladišču proizvaja, dodeluje, odpremlja, ipd. trošarinske izdelke pod ***režimom odloga plačila trošarine***. Le-ta se plača, ko se izdelki iz trošarinskega skladišča sprostijo v uporabo. ***Trošarinsko skladišče*** je posebej označen prostor, namenjen proizvodnji, dodelavi, odpremi, ipd. trošarinskih izdelkov.

Obveznost za obračun trošarine nastane, ko se izdelki sprostijo v uporabo.

Plačila trošarine so oproščeni trošarinski izdelki v prodaji na ladjah in letalih na linijah mednarodnega prometa in v prostocarinskih prodajalnah na mednarodnih letališčih in pristaniščih, izdelki ki jih prinesejo potniki v osebni prtljagi, izdelki v standardnih rezervoarjih motornih vozil, izdelki, namenjeni za službene potrebe diplomatskih predstavnikov in mednarodnih organizacij, itd.

Pri vnosu trošarinskih izdelkov iz druge države EU nastane obveznost za plačilo trošarine če v tujini še niso bili sproščeni v uporabo (t.j. obdavčeni) ali če so v RS namenjeni opravljanju dejavnosti. Ne glede na pravni status osebe, ki vnaša trošarinske izdelke v RS, pa se šteje, da so namenjeni opravljanju dejavnosti, če količine presegajo 800 kosov cigaret, 10 litrov žganja, 90 litrov vina ali 110 litrov piva.

Trošarinski zavezanec mora predložiti carinskemu organu (!) ***obračun trošarine*** za vsako davčno obdobje, ki je koledarski mesec, ne glede na to, ali je v predpisanem obdobju dolžan plačati trošarino ali ne.

3.2.2. Trošarinski izdelki, osnove in stopnje

3.2.2.1. Alkohol in alkoholne pijače

Sem spadajo pivo, vino, druge fermentirane pijače, vmesne pijače ter etilni alkohol. Trošarinska osnova za vina, vmesne in fermentirane pijače je količina izdelka, merjena v hektolitrih, za pivo in etilni alkohol pa prostorninska vsebnost alkohola na en hektoliter.

Trošarina se plačuje v višini:

1. 6,86 EUR za 1% prostorninske vsebnosti alkohola na en hektoliter piva;
2. 0 EUR za en hektoliter mirnega ali penečega vina;
3. 0 EUR za en hektoliter drugih fermentiranih pijač;
4. 62,59 EUR za en hektoliter vmesnih pijač;
5. 694,79 EUR za 100% prostorninske vsebnosti alkohola na en hektoliter etilnega alkohola.

Uporaba etilnega alkohola in alkoholnih pijač je oproščena trošarine, če se uporablja v raziskovalne namene, v zdravstvu ali kot surovina v proizvodnji zdravil, živil, neprehrambenih artiklov ipd.

Mali proizvajalci vina so fizične osebe, ki obdelujejo do 20 hektarjev vinograda ter letno ne pridelajo več kot 100.000 litrov vina. Trošarina se plača za letne proizvedene količine vina (zmanjšane za normirano osebno porabo), trenutno je to 0 EUR za en hektoliter vina.

Mali proizvajalci žganja so fizične osebe, ki uporabljajo kotel za žganjekuho s prostornino nad 40 litrov, ne opravljajo profesionalne dejavnosti žganjekuhe in letno ne proizvedejo več kot 500 litrov žganja. Trošarina se plača v pavšalnem letnem znesku in znaša 12,5 EUR za kotel prostornine 40-100 litrov, ter 25 EUR za večji kotel.

3.2.2.2. Tobačni izdelki

Trošarina se plačuje od cigaret, cigar, cigarilosov in tobaka za kajenje. Trošarinska osnova za tobačne izdelke je 1.000 kosov in drobnoprodajna cena oz. kilogram izdelka.

1. Trošarina od cigaret se plačuje kot specifična trošarina, ki je določena v znesku za 1.000 kosov in kot proporcionalna trošarina določena v odstotku od drobnoprodajne cene cigaret. Skupna trošarina po 1.1. 2004 mora znašati najmanj 57% drobnoprodajne cene zavojčka najbolje prodajanih cigaret v RS, določi pa jo vlada RS na osnovi drobnoprodajnih cen.

2. Trošarina za:
   a) cigare in cigarilose se plačuje v višini 0 EUR za 1000 kosov in 5% od drobnoprodajne cene;
   b) drobno rezan tobak se plačuje v višini 32 EUR v tolarski protivrednosti za en kilogram;
   c) ostali tobak za kajenje se plačuje v višini 20 EUR v tolarski protivrednosti za en kilogram.

3.2.2.3. Energenti in električna energija

Trošarina se plačuje od energentov in električne energije, ki se uporabljajo kot pogonsko gorivo ali za ogrevanje (izjema so biogoriva, šota in biomasa).

Pri električni energiji je trošarinska osnova izražena v megavatnih urah, trošarina pa znaša 1 EUR za eno megavatno uro za neposredno uporabo in 0,5 EUR za poslovno uporabo.

Pri ostalih energentih je trošarinska osnova količina energenta v kilogramih, kubičnih metrih, litrih ali gigajoulih kalorične vrednosti. Trošarine so določene v absolutnih zneskih ter znašajo od 0 EUR za 1.000 kg utekočinjenega naftnega plina za ogrevanje do 413,6371 EUR za 1.000 litrov osvinčenega bencina.

Uporaba energentov je oproščena trošarine, če se uporabljajo kot pogonsko gorivo v letalskem in pomorskem prometu (razen v zasebne namene), za pogon ribiških ladij, v obratih za proizvodnjo električne in toplotne energije ter za nadaljnjo predelavo. Kupec energentov ima pravico do povrnitve 50% trošarine, če se energenti uporabljajo za pogon kmetijske in gozdarske mehanizacije ter za pogon statičnih delovnih strojev, strojev v gradbeništvu, žičnic itd.

Vlada lahko poveča ali zmanjša trošarine za energente in električno energijo do 50% ali določi trošarino za izdelke s stopnjo 0%, ne sme pa določiti trošarin v višinah pod minimalnimi zneski, določene z evropskimi predpisi. Pri alkoholu in alkoholnih pijačah lahko poveča trošarino do 50%, enako pri cigarah, cigarilosih in drobno rezanemu tobaku, ter spremeni razmerje med specifično in proporcionalno trošarino pri cigaretah.

3.3 Davek na motorna vozila

S tem davkom je obdavčen promet:
- novih vozil, ki se dajo prvič v promet oz. se prvič registrirajo na ozemlju RS;
- rabljenih vozil, za katera je obvezna registracija, če od njihovega prometa ni bil obračunan DDV (*obdavčljivih rabljenih vozil*).

Davčni zavezanec je proizvajalec, uvoznik ali oseba, ki pridobi motorno vozilo v drugi državi EU (ne glede na to ali je za takšno vozilo treba plačati DDV po slovenskih predpisih ali ne). Pri prometu obdavčljivih rabljenih vozil pa kupec ali fizična oseba, ki brezplačno ali na podlagi menjave pridobi motorno vozilo, za katero je obvezna registracija, razen če gre za prvi dedni red.

Davek se ne plačuje od vozil, ki se pred prvo registracijo izvozijo ali dobavijo v drugo državo EU, od vozil, ki jih kupijo družine z najmanj tremi mladoletnimi otroki (enkrat v treh letih), od vozil, nabavljenih za prevoz invalidov (enkrat v petih letih), od vozil diplomatskih predstavništev in mednarodnih organizacij v RS, od muzejskih vozil, vozil, ki se začasno uvozijo v RS, športnih vozil, ki se uporabljajo samo na tekmovališčih, od prenosa vozil v primeru statusnega preoblikovanja lastnika vozila.

Davčna osnova pri novem vozilu je prodajna cena novega vozila (brez tega davka in DDV), pri uvozu pa vrednost, določena v skladu s carinskimi predpisi. Davčna lestvica je progresivna (*gre za navadno progresijo!*).

Tabela 2: Davek na motorna vozila

Davčna osnova	**stopnja**
do 4.172,93 EUR	1.0%
nad 4.172,93 EUR do 5.842,10 EUR	1.4%
nad 5.842,10 EUR do 7.511,27 EUR	2.0%
nad 7.511,27 EUR do 10.015,02 EUR	3.5%
nad 10.015,02 EUR do 12.518,78 EUR	5.0%
nad 12.518,78 EUR do 16.691,70 EUR	7.0%
nad 16.691,70 EUR do 20.864,63 EUR	9.0%
nad 20.864,63 EUR do 25.037,56 EUR	11.0%
nad 25.037,56 EUR	13.0%

Pri prometu obdavčljivih rabljenih vozil se plača davek v višini 5% od nakupne cene (če se ta razlikuje od prometne vrednosti, je osnova prometna vrednost, ki jo določi davčni organ).

Za nova vozila, proizvedena po 1.1.2003, katerih uradna specifična emisija ogljikovega dioksida ne presega 110 gramov na prevoženi kilometer, se davek ne plačuje do 31.12.2009.

Davek od motornih vozil se ne všteje v osnovo za DDV. Davek je prihodek proračuna RS.

4. DAVEK OD DOHODKOV PRAVNIH OSEB (ZPOD-2)[3]

4.1. Zavezanec za davek

Davčni zavezanec (*zavezanec*) je pravna oseba domačega prava (davčni rezident) in sicer za davek od vseh dohodkov, ki imajo vir v RS ali izven RS. Zavezanec je tudi pravna oseba tujega prava (davčni nerezident) za davek od dohodkov, ki imajo vir v RS.

Zavezanci za davek niso Republika Slovenija in samoupravne lokalne skupnosti.

Davek je prihodek proračuna RS.

Zavezanci kot so zavodi, društva, ustanove, verske skupnosti, politične stranke, zbornice in sindikati, ki so ustanovljeni za opravljanje nepridobitne dejavnosti so davka oproščeni. Oprostitev pa ne velja v primerih, ko se zavezanci ukvarjajo še s pridobitno dejavnostjo. V tem primeru morajo plačati davek od dohodkov, ki so jih dosegli z opravljanjem pridobitne dejavnosti.

4.2 Davčna osnova in davčna stopnja

Davčna osnova je dobiček, ugotovljen v davčnem izkazu (davčnem obračunu). Davčna osnova se ugotovi tako, da se od ustvarjenih prihodkov (od poslovanja, od financiranja, drugi prihodki) odštejejo davčno priznani odhodki (od poslovanja, od financiranja, drugi odhodki) in davčne olajšave.

Davčna stopnja je 20% od leta 2010 naprej (23% v letu 2007, 22% v letu 2008 in 21% v letu 2009). Investicijski skladi, ki do 30. novembra razdelijo najmanj 90% poslovnega dobička prejšnjega davčnega obdobja, so obdavčeni po stopnji 0%. Po stopnji 0% so obdavčeni tudi pokojninski skladi in zavarovalnice v tistem delu poslovanja, ki se nanaša na izvajanje pokojninskega načrta.

Obdavčenje na podlagi skupinskega davčnega obračuna od leta 2007 ni več možno.

[3] UL, št. 117/06.

4.2.1 Posebnosti pri vključevanju prihodkov in odhodkov v davčno osnovo

Pri poslovanju s ***povezanimi osebami*** (lastniki, njihovimi ožjimi družinskimi člani ter drugimi pravnimi in fizičnimi osebami, ki imajo neposredno ali posredno pravico do udeležbe v upravljanju, nadzoru ali kapitalu zavezanca – eden od glavnih pogojev je 25% delež v kapitalu ali glasovanih pravicah) se pri ***prihodkih in odhodkih*** upoštevajo transferne cene v višini primerljivih tržnih cen, ki bi se dosegle na trgu med nepovezanimi osebami. Prejete in dane obresti med povezanimi osebami pa se priznajo v višini priznane obrestne mere, ki jo predpiše minister za finance.

Pri ugotavljanju davčne osnove se kot odhodek prizna 50% oblikovanih ***rezervacij*** (za dana jamstva, pokojnine, jubilejne nagrade, odpravnine ob upokojitvi), pri bankah, borzno posredniških družbah in zavarovalnicah pa v celoti do višine, ki jo predpisujejo zakon o bančništvu, trgu vrednostnih papirjev oz. zavarovalništvu.

Zavezancu, ki prejema deleže v dobičku od drugih oseb (dividende, deleže dobička d.o.o. ipd.), če je izplačevalec zavezanec za davek, hkrati pa ne sme biti rezident države (izjema so države EU) z ugodnejšim davčnim okoljem (tj. tistih držav, kjer je nominalna stopnja davka od dohodkov pravnih oseb nižja od 12,5%), se prizna izvzem prejetih deležev iz davčne osnove.

Med ***odhodke*** se štejejo samo odhodki, ki so potrebni za pridobitev prihodkov, ki so obdavčeni po ZDDPO-2, torej tisti, ki so neposreden pogoj ali posledica opravljanja dejavnosti, nimajo značaja privatnosti ter so skladni z običajno poslovno prakso. Kot odhodki se ne priznajo:

- odhodki, podobni dividendam, vključno s prikritim izplačilom dobička,
- odhodki za pokrivanje izgub iz prejšnjih let,
- stroški, ki se nanašajo na privatno življenje,
- stroški prisilne izterjave davkov in drugih dajatev,
- kazni, ki jih izreče pristojni organ,
- nekateri davki,
- obresti od nepravočasno plačanih davkov in drugih dajatev, ter od posojil sprejetih od oseb, ki so iz držav z ugodnejšim davčnim okoljem (izjema so države EU),
- podkupnine,
- donacije.

Stroški reprezentance in nadzornega sveta se priznajo v višini 50%.

Odpis terjatev se prizna kot odhodek na podlagi pravnomočnega sklepa sodišča o zaključenem stečajnem postopku ali prisilni poravnavi, v delu, v katerem terjatve niso bile poplačane oziroma niso bile poplačane v celoti. Kot odhodek se odpis terjatev prizna tudi, če zavezanec dokaže, da bi stroški sodnega postopka presegli znesek poplačila terjatev, oz. če dokaže, da je opravil vsa potrebna dejanja za dosego poplačila dolga.

Amortizacija se prizna kot odhodek v obračunanem znesku, vendar največ do zneska, obračunanega z metodo enakomernega časovnega amortiziranja ter najvišjih letnih amortizacijskih stopenj, ki so določene v naslednjih višinah: stavbe 3%, oprema, vozila in mehanizacija 20%, računalniki in računalniška oprema 50%, oprema za raziskovalno dejavnost 33,3%, večletni nasadi 10%, osnovna čreda 20%, druga vlaganja 10%. Presežno obračunana amortizacija se priznava v naslednjih davčnih obdobjih do izteka amortiziranja po davčnih pravilih.

Kot odhodek se priznajo ***plače in nadomestila plač*** za čas odsotnosti z dela v obračunanih zneskih.

Povračila stroškov v zvezi z delom in drugi prejemki zaposlenih (prevoz na delo, stroški prehrane itd.) se priznajo kot odhodek v obračunanih zneskih.

Davčno izgubo (tj. presežek davčnih odhodkov nad prihodki), lahko zavezanec pokriva z zmanjšanjem davčne osnove v naslednjih davčnih obdobjih, pri čemer se najprej zmanjša za izgubo starejšega datuma. Takšno pokrivanje izgube pa ne sme presegati tekoče ugotovljene davčne osnove. Prenos izgube je omejen v primerih, ko se spremeni lastništvo v kapitalu ali glasovalnih pravicah v zavezancu.

4.3 Obdavčitev pri prenosu dejavnosti, zamenjavah kapitalskih deležev, združitvah in delitvah

Zakon o davku od dohodkov pravnih oseb (ZDDPO-2) vsebuje tudi določbe (38. do 53. člen), ki v slovensko zakonodajo uveljavljajo vsebino Direktive EU o skupnem sistemu obdavčitve pri združitvah, delitvah, prenosih sredstev in zamenjavah kapitalskih deležev družb iz različnih držav članic (90/434/EEC). S skupnim sistemom obdavčitve se odpravljajo različne ovire pri preoblikovanji družb, ki so jih vsebovali nacionalni davčni predpisi in to naj bi družbam iz EU pomagalo izboljšali mednarodni konkurenčni položaj. Direktiva in na njej temelječe zakonske rešitve tako omogočajo, da se obdavčitev kapitalskih dobičkov iz

prenesenega premoženja odloži do njihove dejanske realizacije, določena je obdavčitev oz. oprostitev rezervacij, rezerv in izgub pri prenosih ter vrsta drugih davčnih vprašanj, ki se pojavljajo pri statusnem preoblikovanju družb. Družbe, ki se odločajo za opravljanje transakcij na podlagi 38. do 53. člena ZDDPO-2 pa morajo transakcijo priglasiti davčnemu organu.

4.4 Davčne olajšave

4.4.1. Olajšava za vlaganje v raziskave in razvoj

Zavezancu se prizna davčna olajšava v višini 20% investiranega zneska, ki predstavlja vlaganja v raziskave in razvoj oz. celo 30% ali 40% tega zneska, če ima sedež in opravlja dejavnost v manj razvitih regijah države.

Neizkoriščen del davčne olajšave lahko zavezanec pogoji prenese v naslednjih pet davčnih obdobij.

4.4.2. Olajšava za zaposlovanje

Zavezanec je upravičen do davčne olajšave v višini 50–70% izplačanih plač pri njem zaposlenih invalidov, ter do olajšave za praktično delo v strokovnem izobraževanju vajencev, dijakov in študentov v višini do 20% povprečne mesečne plače.

4.4.3. Olajšava za prostovoljno dodatno pokojninsko zavarovanje

Zavezanec, ki financira pokojninski načrt kolektivnega dodatnega pokojninskega zavarovanja lahko uveljavlja zmanjšanje davčne osnove za premije, ki jih plača svojim zaposlenim. Višina olajšave za posameznega delavca je omejena na 24% obveznih prispevkov za pokojninsko in invalidsko zavarovanje tega delavca oz. na 2.390 EUR letno.

4.4.4. Olajšava za donacije

Izplačila za humanitarne, kulturne, znanstvene in podobne namene, izplačana osebam, ki so po predpisih organizirane za opravljanje takih dejavnosti, se priznajo kot olajšava v obliki znižanja davčne osnove, vendar največ v višini 0,3% obdavčenega prihodka zavezanca. Med olajšavo se priznajo tudi izplačila za politične organizacije, vendar največ do zneska treh povprečnih mesečnih plač na zaposlenega.

Vse davčne olajšave skupaj ne smejo presegati višine davčne osnove.

4.5 Obdavčitev dohodkov z davkom po odbitku (tj. »davčnim odtegljajem«)

Pri izplačilu dividend in dividendam podobnih dohodkov, obresti (razen obresti, ki jih izplača RS, zanje jamči RS ter obresti, ki jih banke plačujejo drugim bankam in finančnim institucijam), plačil za uporabo premoženjskih pravic, zakupnin itd., se plača davčni odtegljaj po stopnji 15%. Davčnega odtegljaja se ne plača, kadar gre za izplačila RS, občini, Banki Slovenije, zavezancu rezidentu, ki sporoči svojo davčno številko, ter zavezancu nerezidentu, kadar gre za izplačila njegovi poslovni enoti v RS ter sporoči svojo davčno številko. V primeru da je v mednarodnih pogodba določena nižja stopnja od 15%, se le-ta lahko uporabi.

Davka se ne odtegne v primeru, ko gre za izplačila dividend, obresti ali plačil za uporabo premoženjskih pravic podjetjem, za katere se uporablja skupen sistem obdavčenja, ki velja za odvisne družbe iz različnih držav EU. Osnovni pogoj za to je, da je prejemnik vsaj 24 mesecev lastnik najmanj 10% deleža v kapitalu ali glasovalnih pravicah družbe izplačevalke v primeru dividend in 25% deleža v primeru obresti ali premoženjskih pravic. S tem so v slovensko zakonodajo uvedene vsebine Direktive 90/435/EEC ter Direktive 2003/49/EC katerih namen je odprava dvojne obdavčitve pri izplačilih med povezanimi družbami.

4.6 Obračunavanje in plačevanje davka

Davčna obveznost temelji na načelu samoobdavčitve, po katerem zavezanec sam izračuna davek v davčnem obračunu, ki ga skupaj z izkazom poslovnega izida, bilanco stanja in drugo dokumentacijo predloži do 31. marca za preteklo davčno obdobje, ki je enako koledarskemu letu. Med davčnim obdobjem zavezanec plačuje mesečno ali trimesečno akontacije davka sorazmerno z višino davčne osnove po zadnjem obračunu davka. Davčni zavezanec si lahko izbere tudi davčno obdobje, ki je drugačno od koledarskega leta.

Rezident lahko od davčne obveznosti odšteje znesek tujega davka, ki ga je plačal od dohodkov v tujini, ki so vključeni v njegovo davčno osnovo, vendar največ do višine davka, ki bi bil od teh dohodkov plačan v Sloveniji.

Če zavezanec ne predloži davčnega obračuna ali v postopku nadzora ne zagotovi podatkov za ugotovitev davčne osnove, lahko davčni organ ugotovi davčno osnovo na podlagi ocene prihodkov in odhodkov. Zavezanec lahko doseže znižanje tako ugotovljene davčne osnove samo, če uspe dokazati, da je nižja.

5. PRISPEVKI ZA SOCIALNO VARNOST

5.1 Splošno o sistemu socialnih prispevkov

Prispevki za socialno varnost so po obsegu največja skupina med "veliko trojico" dajatev v našem sistemu javnih financ. Stopnje prispevkov so proporcionalne. Samozaposleni plačujejo zase obe vrsti prispevkov - delodajalčeve in delojemalčeve.

Tabela 5.1: Stopnje prispevkov za socialno varnost 1992-2002 (povprečne stopnje v % od bruto plače)

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002+
Prispevki za PIZ	28,80	30,84	31,00	31,00	26,57	24,35	24,35	24,35	24,35	24,35	24,35
- delojemalci	14,40	15,42	15,50	15,50	15,50	15,50	15,50	15,50	15,50	15,50	15,50
- delodajalci	14,40	15,42	15,50	15,50	11,07	8,85	8,85	8,85	8,85	8,85	8,85
Prispevki za obvezno zdr. zav.	18,15	13,80	12,78	12,70	13,20	13,25	13,25	13,25	13,25	13,25	13,45
- delojemalci	8,69	6,60	6,14	6,10	6,34	6,36	6,36	6,36	6,36	6,36	6,36
- delodajalci	9,46	7,20	6,64	6,60	6,86	6,89	6,89	6,89	6,89	6,89	7,09
Prispevek za zaposlovanje	3,40	3,40	1,25	0,80	0,25	0,20	0,20	0,20	0,20	0,20	0,20
- delojemalci	1,70	1,70	0,63	0,40	0,16	0,14	0,14	0,14	0,14	0,14	0,14
- delodajalci	1,70	1,70	0,63	0,40	0,09	0,06	0,06	0,06	0,06	0,06	0,06
Prispevek za starševsko varstvo	0,00	0,00	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20
- delojemalci	-	-	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
- delodajalci	-	-	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
Skupaj prispevki za socialno varnost	**50,35**	**48,04**	**45,23**	**44,70**	**40,22**	**38,00**	**38,00**	**38,00**	**38,00**	**38,00**	**38,20**
- delojemalci	24,79	23,72	22,36	22,10	22,10	22,10	22,10	22,10	22,10	22,10	22,10
- delodajalci	25,56	24,32	22,86	22,60	18,12	15,90	15,90	15,90	15,90	15,90	16,10

Vir: Zakon o prispevkih za socialno varnost, UL, št. 5/96,97/01

Osnove za plačevanje socialnih prispevkov so (za vse kategorije zavarovancev, razen za kmete, ki so samo zdravstveno zavarovani) enake za pokojninsko in invalidsko zavarovanje, za obvezno zdravstveno zavarovanje, za starševsko varstvo ter prispevek za zaposlovanje.

Prispevki za socialno varnost se plačujejo iz bruto prejemkov. V osnovo za obračunavanje prispevkov se poleg plač oz. zavarovalnih osnov štejejo tudi nadomestila plače za čas odsotnosti z dela. Po Zakonu o prispevkih za socialno varnost (UL, št. 3/98) se je prispevna osnova nekoliko povečala. Tako od leta 1998 zavezanci plačujejo prispevke za socialno varnost tudi od vseh drugih prejemkov iz delovnega razmerja, vključno s stimulacijami in bonitetami. Plačujejo se tudi pri jubilejnih nagradah, odpravninah in solidarnostnih pomočeh ter povračilih stroškov v delu, ki presegajo znesek, ki je določen s predpisom vlade, ter od regresov za letni dopust v delu, ki presega 70% povprečne plače predpreteklega meseca zaposlenih v RS. Prispevki se ne plačujejo od odpravnin, izplačanih zaradi prenehanja delovnega razmerja iz operativnih razlogov ter od premij prostovoljnega dodatnega pokojninskega zavarovanja; slednje je urejeno v pokojninskem zakonu (člena 367 in 368, ZPIZ-2, UL, št. 106/99).

Zavezanci, ki nimajo plač, plačujejo prispevek za obvezno pokojninsko in invalidsko zavarovanje od osnove, določene v skladu s predpisi o obveznem pokojninskem in invalidskem zavarovanju.

Prevladujočo ureditev obveznih socialnih prispevkov pa precej zapleta ureditev prispevkov od ***prenosnih prejemkov***, ki jih zagotavljajo javnofinančne institucije: pokojnine, nadomestila nezaposlenim, bolniške nad 30 dni, starševska nadomestila, invalidska nadomestila, nadomestila za vojaške vaje. Te posebnosti so posledica dejstva, da tovrstni prispevki ne krijejo enakih namenov kot prispevki od plač, po drugi strani pa na dajatve vsaj deloma v praksi vpliva tudi finančna tehnika (nesmiselno bi bilo, če bi finančna institucija, ki izplačuje nadomestila, sama sebi plačevala še delodajalčeve prispevke). Praviloma se pri tovrstnih prejemkih ***ne plačujejo delodajalčevi socialni prispevki***, od nadomestil nezaposlenim tudi ne zaposlovalni prispevki, pri invalidskih nadomestilih pa ne pokojninski prispevki. Od vseh takih prenosnih prejemkov se brez izjem obračuna davek kot akontacija dohodnine. Gotovo pa take manj pomembne posebnosti ureditve prispevkov povzročajo veliko tehničnih težav izplačevalcem pri obračunavanju posameznih prejemkov.

5.2 Prispevki in pravice iz pokojninskega in invalidskega zavarovanja (UL, št. 12/92, UL, št. 106/99)

5.2.1 Uvod

Za zakon o pokojninskem in invalidskem zavarovanju iz leta 1999 (ZPIZ-2) je značilna izjemna netransparentnost, kar je nedvomno posledica pogajanj med socialnimi partnerji in kopice sprejetih amandmajev. Glede na prejšnji zakon iz leta 1992 (ZPIZ-1) seznam osnovnih pravic zavarovancev ni pretrpel bistvenih sprememb, se je pa marsikje spremenil naziv pravice in osnova, od katere se pravica odmerja. ZPIZ-2 večinoma zaostruje pogoje za dodelitev pravic in hkrati zmanjšuje višino teh pravic - predvsem to velja za starostne pokojnine. Poudariti moramo, da bo zaostrovanje pogojev in zmanjševanje višine pravic zelo postopno; v tekstu ta postopnost ni izrecno navedena in prikazana. Izvajanje dela zakona, ki obravnava invalidsko zavarovanje, je bilo "zamrznjeno" do 1. januarja 2003; razlog naj bi bil neprilagojenost zakonodaje na drugih področjih. Velja omeniti, da je ZPIZ-2 občutno omejil možnosti za dodelitev predčasne pokojnine, hkrati pa še vedno omogoča dokupe let, čeprav je cena dokupa bistveno manj ugodna kot nekoč.

Nosilec in izvajalec pokojninskega in invalidskega zavarovanja je Zavod za pokojninsko in invalidsko zavarovanje Slovenije (ZPIZ), ki ima status javnega zavoda.

Sredstva za uresničevanje pravic iz pokojninskega in invalidskega zavarovanja zagotavljajo zavarovanci in delodajalci, za nekatere kategorije zavarovancev pa tudi RS iz svojega proračuna.

5.2.2 Osnove za plačevanje prispevkov za pokojninsko in invalidsko zavarovanje (čl. 207–209, ZPIZ-2, UL, št. 106/99)

1. Za zaposlene v RS je osnova plača oziroma nadomestilo plače; najnižja osnova je minimalna plača.
2. Samozaposleni in kmetje plačujejo prispevke od zavarovalne osnove, ki je določena glede na dosežen dobiček (prihodki manj odhodki). Zavarovalna osnova je odvisna od doseženega dobička, ne more pa biti manjša od minimalne letne plače.
3. Za osebe, ki prejemajo nadomestilo za čas brezposelnosti, je osnova znesek nadomestila.
4. Za osebe, ki so prostovoljno vključena v zavarovanje (to so npr. študentje, osebe, ki skrbijo za otroka ali invalidno osebo, brezposelne osebe, ki niso v evidenci Zavoda za zaposlovanje, ipd.), najnižja osnova ne more biti nižja od zajamčene plače.

5.2.3 Prispevne stopnje

Prispevne stopnje za pokojninsko in invalidsko zavarovanje so:

- zavarovanci po stopnji 15,50%
- delodajalci po stopnji 8,85%

Vsem podjetjem, zavodom in drugim organizacijam za zaposlovanje invalidov in fizičnim osebam, ki so pri njih v delovnem razmerju, se vsi prispevki po tem zakonu obračunajo in odvedejo na poseben račun ter se uporabljajo za materialni razvoj teh podjetij oz. organizacij. Ob tem omenimo, da se prispevek za pokojninsko in invalidsko zavarovanje plačuje tudi na prejemke na podlagi pogodbe o delu, stopnja pa je 6%. Prejemniki teh dohodkov so namreč obvezno zavarovani le za invalidnost, telesno okvaro ali smrt, ki je posledica poškodbe pri delu ali poklicne bolezni (člen 27/2, ZPIZ-2, UL, št. 106/99).

5.2.4 Pravice in pogoji za pridobitev pravic

Z obveznim zavarovanjem se zagotavljajo:

- *pravica do pokojnine* (starostna pokojnina, invalidska pokojnina, vdovska pokojnina, družinska pokojnina, delna pokojnina)

- *pravice iz invalidskega zavarovanja* (pravica do poklicne rehabilitacije, pravica do nadomestila za invalidnost, pravica do premestitve in dela s krajšim delovnim časom od polnega, pravica do drugih nadomestil iz invalidskega zavarovanja, pravica do povrnitve potnih stroškov)
- *dodatne pravice* (pravica do dodatka za pomoč in postrežbo, pravica do invalidnine, varstveni dodatek k pokojnini)
- *druge pravice* (odpravnina, oskrbnina, pravica do dodatka za rekreacijo ali pravica do enkratnega letnega dodatka)

Zavarovanci, ki so prostovoljno vključeni v zavarovanje, se lahko zavarujejo za ožji obseg pravic, ki obsega pravice do starostne, invalidske, vdovske ali družinske pokojnine ter pravico do dodatka za pomoč in postrežbo v primeru slepote.

Za pridobitev pravic do **starostne pokojnine** se zahteva izpolnitev starostnega pogoja in pogoja pokojninske dobe. Pokojninska doba vključuje poleg zavarovalne dobe, tj. dobe, za katero so bili plačani prispevki, tudi posebno dobo, tj. dobo, ki se po zakonu sicer priznava, a za katero niso bili plačani prispevki. Naj ponovno poudarimo, da je dvigovanje starostne meje in zahtevane pokojninske dobe postopno, še zlasti za ženske.

- Zavarovanec pridobi pravico do starostne pokojnine pri starosti 58 let, če je dopolnil 40 let pokojninske dobe (moški) oz. 38 let pokojninske dobe (ženska).
- Zavarovanec pridobi pravico do starostne pokojnine pri starosti 63 let (moški) oz. 61 let (ženska), če je dopolnil 20 let pokojninske dobe.
- Zavarovanec pridobi pravico do starostne pokojnine pri starosti 65 let (moški) oz. 63 let (ženska), če je dopolnil najmanj 15 let zavarovalne dobe.

Minimalni pogoj za dosego starostne pokojnine je 15 let zavarovalne dobe; osebe, ki tega pogoja ne dosegajo, nimajo pravice do starostne pokojnine. *Polna pokojninska doba* je za moške enaka 40 let, za ženske pa 38 let, *polna starost* pa je za moške 63 let, za ženske pa 61 let. Kot bomo videli kasneje, sta oba pojma relevantna pri določanju "bonusov" in "malusov", tj. višjih ali nižjih odmernih stopenj od "normalnih".

Novi zakon uvaja tudi ti. **državno pokojnino**, ki je trajna in jo lahko prejemajo osebe, starejše od 65 let, ki nimajo slovenske ali tuje pokojnine, imajo skromne lastne dohodke, so državljani RS, v obdobju od 15 do 65 leta pa so bivale vsaj 30 let v RS.

Pravice na podlagi invalidnosti so precej raznolike. Najpomembnejša je **pravica do invalidske pokojnine.** V primeru hude invalidnosti (ko zavarovanec ni zmožen opravljati pridobitnega dela) lahko zavarovanec pridobi to pravico brez starostnega pogoja; lažja je oblika invalidnosti, višji je zahtevan starostni prag za dodelitev invalidske pokojnine. Poleg

starostnega pogoja je za dodelitev pravice zahtevan tudi (tako kot pri starostni pokojnini) pogoj pokojninske dobe. Razumljivo je, da je ta pogoj precej milejši kot pri dodelitvi starostne pokojnine: npr. v primeru invalidnosti, ki je nastala zaradi poškodbe pri delu ali poklicne bolezni, se ta pogoj celo ne zahteva, pri drugih vzrokih invalidnosti (poškodbe izven dela ali bolezni) pa se praviloma zahteva, da je ob nastanku invalidnosti zavarovanec imel pokojninsko dobo, ki pokriva vsaj tretjino razdobja od dopolnjenega 20. leta do nastanka invalidnosti. Določene kategorije zavarovancev lahko pridobijo **pravico do poklicne rehabilitacije** (in do določenih nadomestil od trenutka pridobitve te pravice do zaključka rehabilitacije in pričetka dela na novem delovnem mestu). Zavarovanci, ki so končali poklicno rehabilitacijo (in tudi nekateri drugi zavarovanci z lažjo obliko invalidnosti), pridobijo **pravico do premestitve** in hkrati tudi do določenega nadomestila (poleg plače). Zavarovanci, ki imajo lažje oblike invalidnosti, lahko pridobijo **pravico do dela s krajšim delovnim časom** in hkrati pravico do določenega denarnega nadomestila (ki se mu sicer reče delna invalidska pokojnina). **Nadomestilo za invalidnost** danes prejemajo predvsem zavarovanci, ki so nezaposleni in se ne morejo upokojiti.

Vdovska pokojnina je ena številnih novosti; črpanje te pravice je zaščiteno s številnimi varovalkami, da se preprečijo zlorabe. Njena posebnost je, da lahko posameznik poleg te pokojnine prejema tudi starostno pokojnino, vdova npr. lahko prejema starostno pokojnino in odstotek vdovske pokojnine, pri čemer sme skupno izplačilo znašati največ 100% povprečne pokojnine, izplačane v državi v predhodnem letu. **Družinska pokojnina**, namenjena predvsem otrokom umrlega zavarovanca, takšnega kumuliranja ne dovoljuje. **Varstveni dodatek** se dodeli uživalcem pokojnin, v kolikor imajo nizko pokojnino, ki skupaj z drugimi dohodki "ne zadošča za preživetje". Poleg dohodkovnega preizkusa je dodelitev varstvenega dodatka odvisna tudi od premoženjskega preizkusa upokojenca. Pravice do **dodatka za pomoč in postrežbo** so deležni upokojenci, za katere je neogibna stalna pomoč in postrežba drugega. **Invalidnino** (denarno nadomestilo za telesno okvaro) prejemajo osebe, ki so utrpele določene telesne okvare, kar otežuje aktivnost organizma; uživanje te pravice je neodvisno od uživanja drugih pravic iz pokojninskega in invalidskega zavarovanja.

5.2.5 Odmera pokojnin in drugih pravic iz pokojninskega in invalidskega zavarovanja

Starostne pokojnine

Osnova za odmero starostnih, invalidskih in družinskih pokojnin (in tudi nekaterih drugih pravic iz sistema pokojninskega in invalidskega zavarovanja) je *pokojninska osnova*. Pokojninska osnova je mesečno povprečje neto plač, ki jih je zavarovanec dobil v katerihkoli

zaporednih 18. letih (prej 10. letih) zavarovanja po 1.1.1970, ki so zanj najugodnejša. Starostna pokojnina se odmeri od pokojninske osnove:

- Za *zavarovanca* z zavarovalno dobo 15 let 35% pokojninske osnove, nato se za vsako nadaljnje leto poveča za 1,5% (prej 2%); tem stopnjam rečemo stopnje prirasta. Tako npr. lahko zavarovanec s pokojninsko dobo 40 let dobi pokojnino, ki je enaka 72,5% (=35% + 25*1,5%) pokojninske osnove.
- Za *zavarovanko* z zavarovalno dobo 15 let 38% (prej 40%) pokojninske osnove, nato se za vsako nadaljnje leto poveča za 1,5% (prej 2,25%). Zavarovanka s pokojninsko dobo 38 let dobi pokojnino, ki je enaka 72,5% (=38% + 23*1,5%) pokojninske osnove.

V skladu z izjemno netransparentnostjo pokojninskega sistema, se izračunana pokojnina na osnovi zgoraj opisanega postopka dodatno »korigira« s t.i. revalorizacijskimi količniki. To pomeni, da je dejanska končna pokojnina za okoli 20% nižja od pokojnine, določene zgolj na osnovi odmernih stopenj.

Zakon spodbuja kasnejše upokojevanje s tem, da za delovna leta nad polno pokojninsko dobo predpisuje višje stopnje prirasta – to so "bonusi". "Predčasno" upokojevanje pa zakon kaznuje tako, da za osebe, ki nimajo ne polne pokojninske dobe ne polne starosti (izpolnjena morata torej biti oba negativna pogoja) zakon predpisuje "maluse", tj. negativne stopnje prirasta. Potrebno je poudariti, da so višje pozitivne kot tudi negativne stopnje prirasta (od standardnih 1.5%), ki se upoštevajo pri izračunu pokojnine, trajne.

Zakon predpisuje omejitve pri pokojninski osnovi, tako navzgor kot navzdol; na takšen način se zmanjšujejo razlike med najvišjo in najnižjo možno pokojnino. Vrednost najvišje pokojninske osnove je enaka 4 kratniku najnižje pokojninske osnove, vrednost najnižje pokojninske osnove pa je (*de facto*) enaka 64% neto povprečne plače (to je bruto plača manj povprečni davki in prispevki) v RS. Če je izračunana pokojninska osnova posameznika manjša od minimalne pokojninske osnove, se njegova pokojnina izračuna od minimalne pokojninske osnove. Podobno velja tudi za maksimalno pokojninsko osnovo.

Za dva posameznika, ki imata enako pokojninsko dobo, in istočasno vstopata v pokojninski sistem, je maksimalno možno razmerje njunih pokojnin 4:1 (prej 4,8:1).

Državna pokojnina: se odmeri od najnižje pokojninske osnove, znaša pa 33,3% te osnove.

Invalidske pokojnine: se odmerijo od pokojninske osnove. Invalidska pokojnina za primer invalidnosti, ki je posledica nesreče pri delu ali poklicne bolezni, se odmeri tako, kot da bi zavarovanec imel pokojninsko dobo 40 let (moški) oziroma 38 let (ženska). Invalidska pokojnina za invalidnost, ki je posledica bolezni ali poškodbe izven dela, se odmeri na enak

način kot starostna pokojnina, vendar tako izračunana pokojnina ne more biti manjša od 45% pokojninske osnove (moški) oziroma 48% (ženske).

Pri **drugih nadomestilih na podlagi invalidnosti**: to so nadomestilo za čas rehabilitacije, nadomestilo zaradi premestitve, nadomestilo zaradi dela s krajšim delovnim časom in nadomestilo za invalidnost, je osnova izračunana invalidska pokojnina, ki bi jo posameznik dobil, če bi mu bila ta pokojnina odmerjena. Odstotek tako določene osnove se močno razlikuje od nadomestila do nadomestila (npr. pri nadomestilu za čas rehabilitacije je odstotek kar 100%).

Vdovska pokojnina in družinska pokojnina se odmerita od starostne ali invalidske pokojnine, ki jo je prejemal uživalec pokojnine ob smrti. Odmerni odstotek se razlikuje: pri vdovski znaša 70%, pri družinski pa je odvisen od števila odvisnih članov (otroci, starši, ki jih je zavarovanec preživljal ipd.).

Varstveni dodatek: osnova je razlika med pokojnino, ki jo posameznik prejema in najnižjo pokojnino za polno pokojninsko dobo, odmerni odstotek pa narašča z dopolnjeno pokojninsko dobo. Če ima zavarovanec polno pokojninsko dobo, mora očitno biti odmerni odstotek enak 100%.

Dodatek za pomoč in postrežbo in **invalidnina** se odmerita od osnove, ki je enaka najnižji pokojnini za polno pokojninsko dobo; za dodatek za pomoč in postrežbo je odmerni odstotek enak najmanj 70%, pri invalidnini pa je odvisen od stopnje telesne okvare.

5.2.6 Usklajevanje pokojnin in drugih pravic iz sistema pokojninskega in invalidskega zavarovanja

Od julija 2005 (UL, št. 72/2005) se pokojnine usklajujejo na osnovi statističnih podatkov o gibanju povprečnih neto plač.

5.2.7 Dodatno pokojninsko zavarovanje

Dodatno pokojninsko zavarovanje – kot ga pojmuje ZPIZ-2 – je zamišljeno kot drugi steber pokojninskega sistema. Za razliko od osnovnega sistema, tj. prvega stebra, ki je osnovan na sistemu sprotnega prispevnega kritja (PAYG), pa je za dodatno pokojninsko zavarovanje značilno financiranje na osnovi kapitalskega kritja. ZPIZ-2 vzpostavlja podlage za dve vrsti dodatnih zavarovanj: obvezno in prostovoljno. **Obvezno dodatno pokojninsko zavarovanje** vključuje tiste zavarovance, ki delajo na posebno težkih in zdravju škodljivih delovnih mestih; zavezanci za plačilo prispevkov so delodajalci. Po ZPIZ-1, torej v starem

sistemu, so za te skupine zavarovancev delodajalci vplačevali dodatne prispevke, vendar v prvi steber, tj. v obvezni sistem zavarovanja (ti. zavarovalna doba s povečanjem). To novo dodatno pokojninsko zavarovanje bo tem zavezancem omogočalo pridobitev poklicnih pokojnin, ki jih bodo prejemali vse do izpolnitev pogojev za starostno (ali invalidsko) pokojnino, potem pa jo bodo prejemali v zmanjšanem obsegu. Prispevki se stekajo v Sklad obveznega dodatnega zavarovanja, s katerim upravlja Kapitalska družba.

Vlada Republike Slovenije in reprezentativni sindikati javnega sektorja so v prvi polovici 2003 sprejeli odločitev, da se uskladitev plač, predvidena za avgust 2003, trajno preoblikuje v premije kolektivnega dodatnega pokojninskega zavarovanja za javne uslužbence. Tako je nastal zaprti vzajemni pokojninski sklad za javne uslužbence. Upravljavec tega sklada je Kapitalska družba. To pokojninsko zavarovanje je za javne uslužbence obvezno.

Prostovoljno dodatno pokojninsko zavarovanje je dveh oblik: *poklicno* (tj. kolektivno) in *neposredno* (tj. individualno). Kot je znano, so individualne oblike pokojninskega zavarovanja praviloma element tretjega stebra; razlog za to, da ZPIZ-2 *de jure* vključuje individualno dodatno pokojninsko zavarovanje v drugi steber je, da se za vključitev v to obliko zavarovanja vendarle postavljajo določeni pogoji, in sicer, da je zavarovanec vključen v obvezno zavarovanje, oziroma v prvi steber. Za prostovoljno dodatno zavarovanje je značilno, da se financira s premijami (in ne prispevki), ki jih bodo plačevali delodajalci, zavarovanci ali kar oboji. Obseg pravic je odvisen od pravil pokojninskega načrta, ki predpisuje pogoje zavarovanja; natančna vsebina pokojninskega načrta je celo zakonsko predpisana. Sam pokojninski načrt lahko oblikuje delodajalec, zavarovalnica ali banka, pokojninska družba; za upravljavci pokojninskih načrtov pa so zavarovalnice (oziroma banke) ali pokojninske družbe. Če npr. pokojninski načrt predvideva ustanovitev vzajemnega pokojninskega sklada, lahko z njim upravlja le pooblaščena zavarovalnica ali banka. Za spodbujanje tovrstnega pokojninskega zavarovanja ZPIZ-2 ponuja precejšnje davčne olajšave, tako delodajalcem kot zavarovancem. V primeru kolektivnega zavarovanja so prispevki v 2. steber izvzeti iz obdavčevanja pri davku od dobička pravnih oseb, prispevkih za socialno varnost in dohodnini. V primeru individualnega zavarovanja je ugodnost bistveno manjša, ker je prispevek izvzet iz obdavčevanja le pri dohodnini. Seveda, zneski, za katere velja ugodna davčna obravnava so navzgor omejeni in sicer na 24% letnih prispevkov za obvezno pokojninsko in invalidsko zavarovanje oz. na 2.390 EUR letno.

5.3 Prispevki in pravice iz zdravstvenega zavarovanja (UL, št. 9/92, UL, št. 60/02)

5.3.1 Uvod

Z zakonodajo iz leta 1992 je bilo financiranje zdravstva izločeno iz proračuna RS in preneseno na ZZZS (Zavod za zdravstveno zavarovanje Slovenije). Poznamo obvezno in prostovoljno zdravstveno zavarovanje. Nosilec obveznega zavarovanja je ZZZS, prostovoljno pa izvajajo zavarovalnice, bodisi za razliko do polne vrednosti zdravstvenih storitev, za večji obseg pravic, kot jih določa zakon, za višji standard zdravstvenih storitev ali za dodatne pravice, ki niso zajete v obveznem zavarovanju. V Sloveniji prostovoljno zavarovanje trenutno izvajajo zavarovalnice Vzajemna, AdriaticSlovenica in Triglav. Obvezno zavarovanje obsega:

1. *zavarovanje za primer bolezni in poškodbe izven dela;*
2. *zavarovanje za primer poškodbe pri delu in poklicne bolezni.*

Sredstva za obvezno zavarovanje se zagotavljajo s prispevki, ki jih plačujejo ZZZS-ju zavarovanci, delodajalci in drugi z zakonom določeni zavezanci. Prispevki se plačujejo od določenih osnov po proporcionalnih stopnjah, razen za primere, za katere se po zakonu plačujejo prispevki v pavšalnih zneskih. Pavšalni znesek se npr. plačuje od prejemkov na podlagi pogodbe o delu.

5.3.2 Osnove za plačevanje prispevkov

Osnove za plačevanje prispevkov za zdravstveno zavarovanje so enake kot osnove za plačevanje prispevkov za pokojninsko in invalidsko zavarovanje. Edina izjema so kmetje, ki se v obveznem zdravstvenem zavarovanju delijo na dve kategoriji. Kmetje, ki so samo zdravstveno zavarovani (ne pa tudi pokojninsko in invalidsko), plačujejo prispevek za obvezno zdravstveno zavarovanje od katastrskega dohodka kmetijskih in gozdnih zemljišč (52. člen ZZVZZ). Kmetje, ki so pokojninsko in invalidsko zavarovani, pa prispevek za obvezno zdravstveno zavarovanje plačujejo od osnove za pokojninsko in invalidsko zavarovanje, to je od doseženega dobička (209. člen ZPIZ-2); osnova ne sme biti nižja od minimalne plače.

5.3.3 Vrste prispevkov za obvezno zavarovanje, zavezanci in stopnje

I. Za zavarovanje za primer *bolezni in poškodbe izven dela*

1. **Za vse pravice (plačilo zdravstvenih storitev, nadomestilo plače med začasno zadržanostjo od dela, pogrebnina in posmrtnina ter povračilo potnih stroškov)**

- osebe v rednem delovnem razmerju, osebe v rednem delovnem razmerju, ki so bile poslane na delo v tujino, osebe, zaposlene pri tujih in mednarodnih organizacijah, ustanovah in predstavništvih, brezposelne osebe, ki prejemajo denarno nadomestilo oz. denarno pomoč – 6.36%
- delodajalci za osebe v rednem delovnem razmerju, osebe v rednem delovnem razmerju, ki so bile poslane na delo v tujino, osebe, zaposlene pri tujih in mednarodnih organizacijah, ustanovah in predstavništvih, Zavod za zaposlovanje za brezposelne osebe, ki prejemajo denarno nadomestilo oz. denarno pomoč – 6,56%
- osebe, ki samostojno opravljajo gospodarsko ali poklicno dejavnost kot edini ali glavni poklic, osebe, ki so lastniki zasebnih podjetij v RS, če niso zavarovani iz drugega naslova ter osebe s stalnim prebivališčem v RS, zaposlene pri tujem delodajalcu in vrhunski športniki in vrhunski šahisti, če niso drugače zavarovani – 12,92%

2. **Za pravice do zdravstvenih storitev, povračila potnih stroškov, pogrebnino in posmrtnino**

- ZPIZ za upokojene; RS za invalide vojne, vojne žrtve in vojne veterane, za duševno in telesno prizadete ter vojaške obveznike, ki so v civilni službi kot nadomestilu vojaškega roka, vojaške obveznike med služenjem vojaškega roka oz. med usposabljanjem za rezervno sestavo policije; ter osebe z drugimi prihodki – 5,96%

3. **Za pravice do zdravstvenih storitev in povračila potnih stroškov**

- kmetje in člani njihovih gospodarstev, ki opravljajo kmetijsko dejavnost kot edini ali glavni poklic ter osebe s stalnim prebivališčem v Sloveniji, ki prejemajo pokojnino iz tujine – 5,21%
- kmetje, pri katerih je osnova katastrski dohodek od kmetijskih in gozdnih zemljišč – 18,78%

4. **Za pravice do nadomestila plače med začasno zadržanostjo od dela, pogrebnino in posmrtnino**

- kmetje ter člani njihovih gospodarstev, ki opravljajo kmetijsko dejavnost kot edini ali glavni poklic – 1,15%

II. Zavarovanje za primer *poškodbe pri delu in poklicne bolezni*

1. **Za vse pravice (plačilo zdravstvenih storitev, nadomestilo plače med začasno zadržanostjo od dela, pogrebnina in posmrtnina ter povračilo potnih stroškov v zvezi z uveljavljanjem zdravstvenih storitev).**

- delodajalci za delavce v rednem delovnem razmerju, za osebe v rednem delovnem razmerju, ki so bile poslane na delo v tujino, za osebe, zaposlene pri tujih in mednarodnih organizacijah, ustanovah in predstavništvih; samostojni podjetniki, lastniki zasebnih podjetij, vrhunski športniki in šahisti ter kmetje – 0,53%

2. **Za pravice do zdravstvenih storitev, povračila potnih stroškov ter pogrebnino in posmrtnino**

- RS za vojaške obveznike, ki so v civilni službi kot nadomestilu vojaškega roka – 0,18%

Vsem podjetjem, zavodom in drugim organizacijam za zaposlovanje invalidov se *namensko odstopijo* vsi zdravstveni prispevki ter se uporabljajo za materialni razvoj teh podjetij oz. organizacij.

5.3.4 Pravice iz sistema obveznega zdravstvenega zavarovanja

Z obveznim zavarovanjem je zavarovanim osebam plačilo nekaterih ***zdravstvenih storitev*** zagotovljeno ***v celoti*** (UL, št. 9/92, UL, št. 79/94). Za ostala zdravljenja krije ZZZS ***delne stroške***. Odstotke kritja določi ZZZS v soglasju z Vlado RS. Poleg zdravstvenih storitev obvezno zavarovanje krije še stroške ***rehabilitacij in usposabljanja*** po določenih obolenjih.

Zavarovanci imajo tudi pravico do ***nadomestila plač med začasno zadržanostjo od dela (nad 30 dni)*** (za nadomestilo plač med začasno zadržanostjo od dela do 30 dni je zadolžen delodajalec). Osnova za nadomestilo je povprečna mesečna plača zavarovanca. Nadomestilo znaša:

- 100% od osnove ob zadržanosti od dela zaradi poklicne bolezni, poškodbe pri delu, presaditve organov in tkiva v korist druge osebe, posledic dajanja krvi,
- 90% osnove ob zadržanosti od dela zaradi bolezni,
- 80% osnove ob zadržanosti od dela zaradi poškodb izven dela, nege družinskega člana in spremstva, ki ga odredi zdravnik.

Seznam ***zdravil***, katerih stroški se krijejo iz obveznega zavarovanja, je določen z razvrstitvijo zdravil na liste, ki jih določi posebna Komisija za razporeditev zdravil na liste, ki jo je ustanovil ZZZS. Zavarovane osebe imajo pri uresničevanju pravic do zdravstvenih storitev pravico do povračila ***potnih stroškov,*** če zaradi zdravljenja in diagnostike večkrat na mesec potujejo v drug kraj. Ob smrti zavarovane osebe pripada ***pogrebnina*** osebi, ki je poskrbela za pogreb. Družinski člani zavarovanca, ki jih je ta preživljal do svoje smrti, imajo pravico do ***posmrtnine*** kot enkratne denarne pomoči ob smrti zavarovanca. Višina pogrebnine in posmrtnine je odvisna od dohodka umrle zavarovane osebe v koledarskem letu pred letom smrti. Pogrebnina znaša 40 oz. 60%, posmrtnina pa 10 oz. 25% mesečne bruto plače v RS za obdobje januar-september preteklega leta.

Širok obseg pravic povzroča stalna neskladja med zdravstvenimi potrebami in željami prebivalstva ter potrebnimi finančnimi sredstvi za izvajanje zdravstvenih programov. Odprava teh neskladij se rešuje s postopnim krčenjem obsega pravic, uvajanjem prostovoljnega zavarovanja, prerazvrščanjem zdravil med listami ter nenazadnje višanjem prispevne stopnje za obvezno zdravstveno zavarovanje.

5.3.5 Prostovoljno zdravstveno zavarovanje

Namen prostovoljnega zavarovanja je možnost, da si zavarovanci proti ustreznemu plačilu iz lastnih sredstev omogočijo boljši standard, boljši dostop do zdravstvenih storitev ter boljše pogoje pri uveljavljanju zahtev in potreb po zdravstvenih storitvah in socialni varnosti, ki jih obvezno zdravstveno zavarovanje s sredstvi javnih financ ne zagotavlja. V Sloveniji je trenutno prostovoljno zavarovano več kot poldrugi milijon prebivalcev (pri zavarovalnicah Vzajemna, AdriaticSlovenica in Triglav). Večina jih je zavarovanih za programe doplačil, nekaj pa tudi za nadstandardne storitve. Znesek premije je "skoraj" enak za vse zavarovance. "Skoraj" pravimo zato, ker je višina premije dejansko odvisna od vstopne starosti zavarovanca. Zakon o spremembah in dopolnitvah zakona o zdravstvenem varstvu in zdravstvenem zavarovanju (UL, št. 29/98) dopušča, da se premija oblikuje glede na starost, spol, tablice smrtnosti in bolezenske tablice itd., vendar v praksi takšna "razdelava" premije še ni realizirana. Premija je obravnavana kot boniteta, če jo plača delodajalec (in obdavčena z delodajalčevimi in delojemalčevimi prispevki za socialno varnost ter dohodnino).

5.4 Družinski prejemki in starševsko varstvo (UL, št. 97/01, UL, št. 76/03)

5.4.1 Vrste prispevkov, prispevne stopnje in osnove

Sredstva za starševsko varstvo se zbirajo s prispevki delodajalca in zavarovanca in se vplačujejo v proračun RS. Plačujejo ga:

- delodajalci – 0,10%
- zavarovanci, to so vsi zaposleni na območju RS, izvoljeni ali imenovani nosilci oblasti v RS, tujci, ki so zaposleni v RS pri mednarodnih organizacijah ter konzularnih in diplomatskih predstavništvih, samozaposleni, kmetje, vrhunski športniki in šahisti, brezposelni, osebe, ki so na prestajanju kazni, osebe, ki prejemajo nadomestilo med začasno zadržanostjo od dela od ZZZS ter osebe, ki prejemajo starševsko nadomestilo – 0,10%

5.4.2 Pravice

Pravice se delijo v dva sklopa:
1. pravice do starševskega varstva
2. pravice do družinskih prejemkov

5.4.2.1 Pravice do starševskega varstva

Pravice iz zavarovanja za starševsko varstvo so:
1. starševski dopust
2. starševsko nadomestilo
3. pravica iz naslova krajšega delovnega časa

1. Pravico do starševskega dopusta imajo starši, ki so zavarovani. Starševski dopust ima več oblik:
   - ***porodniški dopust***, ki ga izrabi mati v strnjenem nizu v obliki polne odsotnosti z dela in traja 105 dni. Mati mora nastopiti porodniški dopust 28 dni, lahko pa 42 dni pred predvidenim datumom poroda;
   - ***očetovski dopust*** v trajanju 90 dni, ki ga izkoristi oče v času porodniškega dopusta matere najmanj v trajanju 15 dni v obliki polne odsotnosti z dela. Ostalih 75 dni lahko oče koristi v obliki polne odsotnosti z dela do 8. leta starosti otroka. Očetovski dopust se v Sloveniji uvaja postopoma. Začel se je uvajati z letom 2003, ko se je priznala pravica do očetovskega dopusta v trajanju 15 dni, z letom 2005 pa je očetovski dopust že dosegel trajanje polnih 90 dni.

- ***dopust za nego in varstvo otroka*** lahko izkoristi eden od staršev v trajanju 260 dni neposredno po preteku porodniškega dopusta. Starša se pisno dogovorita o izrabi dopusta za nego in varstvo otroka 30 dni pred potekom porodniškega dopusta ter dogovor predložita pristojnemu Centru za socialno delo skupaj z zahtevo za uveljavljanje pravice. Dopust se izrabi v strnjenem nizu v obliki polne ali delne odsotnosti z dela.
- ***posvojiteljski dopust*** izrabi posvojitelj oz. oseba, ki ji je otrok zaupan v vzgojo in varstvo z namenom posvojitve in sicer v trajanju 150 dni za otroka, starega od 1. do 4. let ter v trajanju 120 dni za otroka, starega od 4. do 10. let.

2. Pravico do starševskega nadomestila imajo tiste osebe, ki imajo pravico do starševskega dopusta in so bile pred dnevom nastopa posamezne vrste starševskega dopusta zavarovane za starševsko varstvo. Pravico do starševskega nadomestila imajo prav tako osebe, ki nimajo pravice do starševskega dopusta, so pa bile zavarovane najmanj 12 mesecev v zadnjih treh letih pred nastopom posamezne vrste starševskega dopusta. Pravica od starševskega nadomestila obsega:
   - ***porodniško nadomestilo*** v času porodniškega dopusta,
   - ***očetovsko nadomestilo*** v času očetovskega dopusta,
   - ***nadomestilo za nego in varstvo otroka*** v času dopusta za nego in varstvo otroka ter
   - ***posvojiteljsko nadomestilo*** v času posvojiteljskega dopusta.

   Osnova za posamezno vrsto starševskega nadomestila je povprečna mesečna plača oz. povprečna osnova, od katere so bili obračunani prispevki za starševsko varstvo v zadnjih 12 mesecih pred nastopom starševskega dopusta. Denarno nadomestilo je povprečna mesečna plača upravičenca oz. povprečna osnova, od katere je upravičenec plačeval prispevek za porodniško varstvo v zadnjih 12 mesecih. Če je upravičenec prejemal plačo za krajše obdobje, se mu za manjkajoče mesece kot osnova upošteva 55% minimalne plače. Osnova ne more biti višja od dvainpolkratnika povprečne mesečne plače v RS, razen za porodniško nadomestilo ter ne more biti nižja od 55% minimalne plače. Starševsko nadomestilo znaša 100% osnove, razen za očetovsko nadomestilo, ki znaša 100% osnove v trajanju 15 dni, za ostalih 75 dni pa očetu RS zagotavlja plačilo prispevkov za socialno varnost od minimalne plače. Starševska nadomestila se usklajujejo z rastjo izhodiščne plače za negospodarske dejavnosti.

3. Pravica iz naslova krajšega delovnega časa: eden od staršev, ki neguje in varuje otroka do tretjega leta starosti, ima pravico delati krajši delovni čas, pri čemer mora delovni čas obsegati najmanj polovično tedensko delovno obveznost.

5.4.2.2 Pravice do družinskih prejemkov

Družinski prejemki so denarni prejemki, ki obsegajo:
1. starševski dodatek,
2. pomoč ob rojstvu otroka,
3. otroški dodatek,
4. dodatek za veliko družino,
5. dodatek za nego otroka,
6. delno plačilo za izgubljeni dohodek.

Vsaka mati, ki je državljanka RS, ima stalno prebivališče v RS ter ima otroka, ki je državljan RS ter ni upravičena do starševskega nadomestila, je upravičena do **starševskega dodatka,** 365 dni od rojstva otroka. **Pomoč ob rojstvu otroka** je enkratni denarni prejemek, namenjen nakupu opreme za novorojenca. Pomoč se lahko zagotovi tudi v obliki zavitka, ki vsebuje osnovno opremo za novorojenca. Pomoč pripada otroku, katerega oče ali mati imata stalno prebivališče v RS. Z **otroškim dodatkom** se staršem oz. otroku zagotovi dopolnilni prejemek za preživljanje, vzgojo in izobraževanje, njegova višina pa je odvisna od uvrstitve družine v dohodkovni razred in števila otrok. Za uvrstitev družine v dohodkovni razred se upošteva povprečni mesečni dohodek na družinskega člana v preteklem koledarskem letu. Med dohodke se vštevajo vsi bruto dohodki, ki so vir dohodnine, transferni dohodki in vsi drugi dohodki, razen namenskih dodatkov, kot so npr. pomoč ob rojstvu otroka, otroški dodatek, dodatek za nego otroka, dodatka za pomoč in postrežbo itd. Pravico do otroškega dodatka ima otrok do dopolnjenega 18. leta starosti oz. do 26. leta, dokler ima otrok status učenca, dijaka, vajenca ali študenta na dodiplomskem študiju. Pravico do otroškega dodatka ima eden od staršev, če je državljan RS (ali če ni državljan RS ob pogoju vzajemnosti) in če ima otrok prebivališče v RS. **Dodatek za veliko družino** je letni enkratni prejemek, namenjen družini z več otroki. Velika družina je družina, ki ima tri ali več otrok. Pravico do dodatka ima eden od staršev, če so starši in otroci državljani RS in imajo stalno prebivališče v RS. Pravica do dodatka preneha, ko najstarejši otrok dopolni z zakonom predpisano starost oz. preneha s šolanjem. **Dodatek za nego otroka** je denarni dodatek za otroka, ki potrebuje posebno nego in varstvo in je namenjen kritju povečanih življenjskih stroškov, ki jih ima družina z nego in varstvom takega otroka. Pravico do dodatka ima eden od staršev, če je otrok državljan RS in ima v RS stalno prebivališče**. Delno plačilo za izgubljeni dohodek** je osebni prejemek, ki ga prejme eden od staršev, kadar prekine delovno razmerje ali začne delati krajši delovni čas zaradi nege in varstva otroka s težko motnjo v duševnem razvoju. Mesečna višina delnega plačila je minimalna plača. Pravica se uveljavlja na podlagi mnenja zdravniške komisije.

Družinski prejemki se usklajujejo enkrat letno (januarja) z indeksom rasti cen življenjskih potrebščin.

5.5 Prispevki in pravice iz zavarovanja za primer brezposelnosti (UL, št. 71/93, UL, št. 67/02)

5.5.1 Vrste prispevkov, prispevne stopnje in osnove

Sredstva za zavarovanje za brezposelnost (denarna nadomestila in denarna pomoč za čas brezposelnosti) razporeja Republiški zavod za zaposlovanje. Sredstva, s katerimi razpolaga, se zagotavljajo delno s prispevki (delavcev iz plač, delodajalcev od izplačanih plač) in se zbirajo neposredno v proračunu RS, delno pa iz ostalih virov republiškega proračuna.

Prispevek za zaposlovanje plačujejo:
- zavarovanci po stopnji 0,14%,
- delodajalci po stopnji 0,06%.

5.5.2 Pravice

Zakonsko pravico do ***denarnega nadomestila*** pridobijo brezposelne osebe, ki so bile pred nastankom brezposelnosti zavarovane za primer brezposelnosti in za katere ni na voljo ustrezne zaposlitve, ki so zaposlitev izgubile brez lastne volje ali krivde ter so bile pred prenehanjem delovnega razmerja v delovnem razmerju pri enem ali več delodajalcih vsaj dvanajst mesecev v zadnjih osemnajstih mesecih. Oseba se v roku 30 dni po izteku delovnega razmerja prijavi pri zavodu za zaposlovanje in uveljavlja pravico do nadomestila.

Status brezposelne osebe (ki je pogoj za uveljavljanje pravice do nadomestila) ima oseba, če izpolnjuje naslednje pogoje:
- ni v delovnem razmerju,
- ni samozaposlena oseba ali lastnik ali solastnik podjetja, v katerem je ustvaril dobiček, ki bi presegal znesek zajamčenega nadomestila plače,
- ni lastnik, zakupnik, najemnik ali drug uporabnik kmetijskega ali gozdnega zemljišča s katastrskim dohodkom, ki bi dosegel višino, ki je določena kot podlaga za vključitev v obvezno pokojninsko in invalidsko zavarovanje,
- ni upokojenec, študent, dijak, vajenec in je zmožen za delo, prijavljen pri zavodu, na razpolago za zaposlitev in aktivni iskalec zaposlitve.

Čas izplačevanja denarnega nadomestila je odvisen od dosežene delovne dobe, izplačuje pa se *od treh mesecev do dveh let*, razen v izjemnih primerih, ko upravičencu do denarnega nadomestila ob izteku roka za izplačevanje do izpolnitve pogojev za upokojitev manjkajo največ tri leta. V tem primeru brezposelni osebi zavod plačuje prispevek za pokojninsko in invalidsko zavarovanje do izpolnitve pogojev za upokojitev - zakon ne predvideva več

možnosti podaljšanja izplačevanja denarnega nadomestila, prav tako ne sofinanciranja dokupa manjkajoče zavarovalne dobe. Osnovo za odmero denarnega nadomestila predstavlja povprečna mesečna plača brezposelne osebe v zadnjih dvanajstih mesecih pred nastankom brezposelnosti. Prve tri mesece denarno nadomestilo znaša 70%, v naslednjih mesecih pa 60% od tako določene osnove. Zakon določa zgornjo (300% zajamčene plače) in spodnjo mejo (100% zajamčene plače) višine nadomestila. Od odmerjenega nadomestila se obračuna prispevke za socialno varnost po stopnjah, določenih s predpisi ter akontacija dohodnine.

Prejemnik denarnega nadomestila ima zakonsko možnost, da na podlagi pogodbe o zaposlitvi (ne več kot polovico delovnega časa), avtorske pogodbe ali kako drugače pridobiva dodatne dohodke. Denarno nadomestilo se v takih primerih zniža za 50% dodatnega dohodka pri izplačilu denarnega nadomestila v naslednjem mesecu.

Brezposelna oseba mora v času prejemanja denarnega nadomestila sprejeti zaposlitev na delovnem mestu, ki ustreza njeni izobrazbi, znanju in zmožnostim ter se vključevati v programe aktivne politike zaposlovanja, v katere je napotena s strani zavoda za zaposlovanje. Po izteku pravice do denarnega nadomestila lahko brezposelna oseba uveljavi pravico do ***denarne socialne pomoči*** (UL, št. 79/2006). Leta 2006 je bila ukinjena posebna socialnovarstvena pravica, namenjena samo brezposelnim osebam, ki so izčrpali pravico do denarnega nadomestila (to je bila denarna pomoč). Tako so sedaj brezposelni, ko se jim izteče pravica do denarnega nadomestila upravičeni (če ne dosegajo dohodkovnega cenzusa) do že omenjene denarne socialne pomoči, ki je »splošna« socialnovarstvena pravica. Za prejemanje te pomoči morajo osebe poleg dohodkovnega cenzusa izpolnjevati dodatni pogoj – to je pogoj aktivnega iskanja dela.

5.6 Socialno-varstveni prejemki

Za socialno-varstvene prejemke je značilno, da je upravičenje osnovano na nekem dohodkovnem cenzusu. Nekatere od teh pravic so namenjene povsem določeni skupini. Tako sta varstveni dodatek in državna pokojnina namenjena upokojencem, denarna pomoč za brezposelne je namenjena nezaposlenim, republiška štipendija pa je namenjena dijakom in študentom. Te pravice se izplačujejo iz državnega proračuna oziroma proračunov zavodov za socialno zavarovanje. Poleg zgoraj navedenih denarnih prejemkov, ki se izplačujejo posameznikom, se določene storitve subvencionirajo in s tem se znižuje cena storitve za ciljno skupino. Tako se npr. subvencionira prehrana v osnovnih in srednjih šolah, in ta subvencija je namenjena dijakom iz revnejših družin.

Tudi lokalna raven izplačuje subvencije organizacijam, ki zagotavljajo določene storitve. Med drugim so otroški vrtci deležni pomembnih subvencij, prav tako se na lokalni ravni subvencionira dolgotrajna nega in oskrba, bodisi institucionalna bodisi na domu. Na lokalni ravni se subvencionirajo tudi najemniki v neprofitnih stanovanjih.

Zgoraj omenjene pravice, ki sicer imajo značaj socialno-varstvenih pravic, niso izrecno navedene v zakonu o socialnem varstvu, temveč so podrobneje opredeljene v posameznih »področnih zakonih (npr. zakon o pokojninskem in invalidskem zavarovanju itd). Zakon o socialnem varstvu (Uradni list 26/2001) posebej našteva oziroma določa štiri vrste denarnih socialnih pomoči:

- denarna socialna pomoč za obdobje; ta pravica se dodeli za določen čas; maksimalno trajanje je eno leto,
- denarna socialna pomoč – trajna; se dodeli upravičencu nad 60 let starosti in tistemu, ki je trajno nezmožen za delo in brez vsakršnih dohodkov in nima koga, ki bi ga bil po zakonu dolžan preživljati in živi doma,
- izredna denarna socialna pomoč – za obdobje; se dodeli v izrednih okoliščinah tudi posameznikom, ki presegajo dohodkovni cenzus (obdobje znaša do dveh mesecev),
- izredna denarna socialna pomoč – enkratna; se dodeli v enkratnem znesku.

Denarna socialna pomoč (za obdobje in trajna) je določena kot razlika med minimalnih dohodkom (vrednost se določa enkrat letno) in dohodkom družine/posameznika. Od teh štirih vrst socialnih pomoči je daleč najpomembnejša denarna socialna pomoč za obdobje, ki predstavlja skoraj 90% vseh izdatkov za denarne socialne pomoči.

5.7 Usklajevanje socialnih transferjev

Sistem socialnih transferjev je – kot smo videli – zelo razvejan in, podobno kot pokojninski sistem, izjemno netransparenten. Pomemben korak k ureditvi tega področja je bil dosežen s sprejetjem zakona o usklajevanju transferjev posameznikom in gospodinjstvom (UL, št. 114/2006), ki je določil enotni način usklajevanja teh transferjev, in sicer z rastjo cen življenjskih potrebščin. Usklajevanje se opravi enkrat letno, v mesecu januarju, upošteva pa se rast teh cen v obdobju januar-december preteklega leta.

6. DAVKI NA PLAČILNE LISTE

6.1 Davek na izplačane plače (ZPOD)[4]

Davek plačujejo pravne in fizične osebe, ki izplačujejo plače in so po posebnih zakonih zavezanci za plačevanje prispevkov za pokojninsko in invalidsko zavarovanje, obvezno zdravstveno zavarovanje, starševsko varstvo in za zaposlovanje. Davčna osnova so bruto plače. Davka ne plačujejo invalidska podjetja in tuja diplomatska predstavništva ter konzulati. Davek je prihodek proračuna RS. Uveden je bil leta 1996 zaradi primanjkljaja, ki je nastal kot posledica znižanja delodajalčevih prispevkov na bruto plače ter istočasno kot poskus izenačitve položaja delovno intenzivnih panog, kjer so plače nižje, z ostalimi panogami. V novembru 2005 sprejeti zakon (UL 108/2005) davek v letih do 2009 postopoma ukinja. Stopnje in osnova davka so:

Osnova (EUR)	2005	2006	2007	2008
do 688,53	0%	0%	0%	0%
688,53 - 1.669,17	3,8%	3,0	2,3	1,1
1.669,17 - 3.129,69	7,8%	6,3	4,7	2,3
nad 3.129,69	14,8%	11,8	8,9	4,4

S 1.1.2009 obveznost plačevanja davka na izplačane plače preneha.

6.2 Posebni davek na določene prejemke (ZPOD)

Posebni davek na določene prejemke (za opravljene storitve na podlagi pogodbe o delu) je bil uveden leta 1993 zaradi prejšnje prenizke skupne obdavčitve takih prejemkov v primerjavi s plačami iz rednega delovnega razmerja. Zaradi prenizke obdavčitve je bila taka oblika zaposlitve pred uvedbo tega davka zelo razširjena.

Davek plačujejo pravne osebe in zasebniki, ki opravljajo dejavnost in izplačujejo določene prejemke *na podlagi pogodbe o delu*. Osnova za obračun in plačilo davka je vsako posamezno *bruto izplačilo* fizični osebi za opravljeno storitev na podlagi pogodbe o delu ter vsa *povračila stroškov*, ki jih je fizična oseba prejela v zvezi z opravljanjem storitve. Za izplačila za opravljeno storitev se ne štejejo (a) plačila za uporabo avtorskega dela, (b) plačila za delo učencem in študentom, (c) plačila za nego invalidov, (d) plačila fizičnim osebam, ki opravljajo funkcije na podlagi imenovanja državnega ali upravnega organa, (e) do določene višine nagrade in povračila stroškov fizični osebi, ki prostovoljno sodeluje v

[4] Zbirka predpisov o dajatvah.

ljubiteljskih, humanitarnih in podobnih dejavnostih ter v strankarskih ali sindikalnih dejavnostih, (f) izplačila fizični osebi za spravilo pridelkov ali ulova v okviru sezonskih del, (g) izplačila pri sezonskem tradicionalnem izlovu cipljev, (h) izplačila fizični osebi, ki opravi storitev na podlagi napotitve upravnega organa ter (i) izplačila fizični osebi na podlagi poziva na določeno opravilo s strani državnega organa.

Davek se plačuje po stopnji 25% in je prihodek proračuna RS.

7. PREMOŽENJSKI IN DRUGI DAVKI

7.1 Davek od premoženja (UL, št. 36/88)

V Sloveniji z davkom od premoženja obdavčujemo fizične osebe, ki posedujejo stavbe, dele stavb, stanovanja in garaže ali prostore za počitek ali rekreacijo (tj. vikende).

Davek od premoženja je prihodek lokalnih skupnosti. Ker je obdavčitev premoženja v Sloveniji zastarela in so premoženjski davki relativno manj izdatni kot v drugih evropskih državah, se pri teh davkih pričakuje skorajšnja reforma.

Davčno osnovo predstavlja vrednost stavbe oziroma prostora za počitek ali rekreacijo. Določi jo upravni organ po sistemu točkovanja, vrednost točke pa se letno prilagaja rasti cen življenjskih potrebščin.

Pri stanovanjskih površinah velja oprostitev za prvih 160 kvadratnih metrov, vendar le pod pogojem, da v njej prebiva lastnik ali njegovi ožji družinski člani. Plačila davka na stavbe so oproščena kmetijska poslopja, poslovni prostori, ki jih lastnik uporablja za opravljanje dejavnosti in stavbe, ki so razglašene za kulturne ali zgodovinske spomenike. Nove stavbe so plačila oproščene prvih 10 let, oprostitve je mogoče uveljavljati tudi ob izdatnejših prenovah itd. Davčne stopnje za stavbe so v razponu od 0,10 % do 1 %.

Nekoliko bolj so obdavčeni vikendi, saj se pri njih obdavči celotna površina, davčne stopnje pa so od 0,15 % do 1,5 %, ni pa mogoče uveljavljati 10-letne olajšava za nove stavbe in olajšave ob večjih prenovah. Za poslovne prostore se davčna stopnja giblje med 0,15 % in 1,25 % (pri poslovnih stavbah, ki jih lastnik ne uporablja za opravljanje dejavnosti in jih tudi ne oddaja v najem, se stopnje povišajo za 50%).

7.2 Davek na vodna plovila (UL, št. 117/06)

Davek na vodna plovila se plačuje od plovil daljših od petih metrov, ki so vpisana v evidencah plovil oziroma izpolnjujejo tehnične pogoje za vpis v evidence plovil (tudi plovila, registrirana v tujini). Davčni prihodek pripada občini stalnega ali začasnega prebivališča oziroma sedeža zavezanca.

Davčni zavezanec je lastnik oziroma solastnik plovila, pri plovilu, ki je predmet pogodbe o lizingu, pa je davčni zavezanec lahko tudi uporabnik, če lastnik in uporabnik plovila o tem skupno obvestita davčni urad.

A Primer in Game Theory

Robert Gibbons

FT Prentice Hall
FINANCIAL TIMES

An Imprint of Pearson Education
Harlow, England • London • New York • Boston • San Francisco • Toronto
Sydney • Tokyo • Singapore • Hong Kong • Seoul • Taipei • New Delhi
Cape Town • Madrid • Mexico City • Amsterdam • Munich • Paris • Milan

Chapter 1

Static Games of Complete Information

In this chapter we consider games of the following simple form: first the players simultaneously choose actions; then the players receive payoffs that depend on the combination of actions just chosen. Within the class of such static (or simultaneous-move) games, we restrict attention to games of *complete information*. That is, each player's payoff function (the function that determines the player's payoff from the combination of actions chosen by the players) is common knowledge among all the players. We consider dynamic (or sequential-move) games in Chapters 2 and 4, and games of incomplete information (games in which some player is uncertain about another player's payoff function—as in an auction where each bidder's willingness to pay for the good being sold is unknown to the other bidders) in Chapters 3 and 4.

In Section 1.1 we take a first pass at the two basic issues in game theory: how to describe a game and how to solve the resulting game-theoretic problem. We develop the tools we will use in analyzing static games of complete information, and also the foundations of the theory we will use to analyze richer games in later chapters. We define the *normal-form representation* of a game and the notion of a *strictly dominated strategy*. We show that some games can be solved by applying the idea that rational players do not play strictly dominated strategies, but also that in other games this approach produces a very imprecise prediction about the play of the game (sometimes as imprecise as "anything could

Example 1.7 Consider a firm that will liquidate one period hence at time t = 1. There are no taxes and the firm can invest $30 in a risky venture at t = 0 using retained earnings. If the investment is not made, shareholders get a dividend of $100 at t = 0. The firm's debt requires a payment of $100 at t = 1, and its investment choices are described in Table 1.3.

TABLE 1.3 Payoffs Related to Different Investment Opportunities

	State of Nature	
Strategy	**Boom (with probability 0.5)**	**Bust (with probability 0.5)**
Total firm value at $t = 1$ if no investment made and $100 dividend paid at $t = 0$	$110	$70
Total firm value at $t = 1$ if $30 investment made and $70 dividend paid at $t = 0$	$200	$ 5

For simplicity, assume that the discount rate is zero. What should the firm do?

Solution To analyze this problem, first compute the net present value (NPV) of each choice for the firm as a whole. If it does not invest, then its expected value is 0.5(110) + 0.5(70) = $90. Add to this the $100 dividend paid at t = 0 and we get a total firm value of $190. If it does invest, then its expected value is 0.5(200) + 0.5(5) = $102.5. Add to this the $70 dividend paid at t = 0 and we get a total firm value of $172.5. Since total firm value is lower with the investment than without, the project has negative NPV. The apparent choice should be to reject the investment.

Hold it for a minute, though! This decision rule is the right one only if you want to maximize total firm value. But remember that your goal is to maximize the wealth of the shareholders. If there is no investment, the shareholders get $100 dividend plus $10 ($110 debt payment) in the boom state and nothing in bust state (limited liability, which stipulates that the liability of the shareholder does not extend beyond the assets of the firm, means that the bondholders get $70 and the shareholders get 100 + 0.5(10) = $105. On the other hand, if the project is accepted, they get $100 in the boom state and nothing in the bust state. Thus the value of this strategy to the shareholder is 70 + 0.5(100) = $120. Clearly, the shareholders want you to invest in the project. Thus, a project with negative NPV for the firm as a whole may be chosen in the best interest of the shareholder.

This example illustrates a moral hazard faced by bondholders. The firm, acting in the interest of the shareholders, has an incentive to undertake investments that benefit the shareholders at the expense of creditors. In this example, the expected payoff to the bondholders is 0.5(100) + 0.5(70) = $85 if the firm does not invest in the risky project and 0.5(100) + 0.5(5) = $52.50 if the firm invests in the risky project. Thus, by investing in the risky project, the shareholders reduce the wealth of the bondholders by $32.50. The shareholders themselves gain $15, so that there is a net decline in total firm value of $17.50. This is the aggregate loss due to moral hazard.

In this example, we assumed that the manager acted in the best interest of the shareholders. However, that is a questionable assumption too.[12] As an agent of the shareholders, the managers can do many things that may not be in the interest of the shareholders. For example, by inflating expenses, management can divert earnings from shareholders to management. Likewise, managers can discourage takeovers and thereby entrench themselves at the possible expense of shareholders. Managers may also select myopic and low-risk investment projects with a view toward protecting their positions and reputations.

You may have noticed that a critical assumption made in these examples is that the principal (the insurance company, the bondholders, or the shareholders) is unable to completely control the agent's behavior. If it were possible to costlessly observe the agent's actions, there would be no moral hazard. If the insurance company could precisely observe the insured, it would simply prohibit all actions detrimental to the car. It is because final outcomes do not unambiguously reveal the actions that may have influenced them that such proscriptions cannot be effectively written into contracts. Thus, for moral hazard to arise, it must be that: (i) the agent's actions (that affect the final outcome) cannot be costlessly observed by the principal, and (ii) there is some noise (exogenous uncertainty) that masks the agent's action in the final outcome.

Of course, the principal anticipates the agent's behavior. Thus, the principal attempts to design a contract that aligns the agent's incentives with her own. Deductibles and other coinsurance provisions in insurance contracts serve this purpose. Bondholders address moral hazard by limiting the firm's debt (the higher the debt/equity ratio, the greater is the inclination of shareholders to choose risky projects), by requiring collateral,[13] and by including in the debt contract covenants that restrict the borrower's actions. The interests of managers are aligned with the interests of shareholders through compensation contracts that include stock and stock options.

Another way to address moral hazard is to contract with the agent over extended time periods. Because of the possibility of reputational consideration, the agent may restrain self-interested behavior that is to the principal's detriment.[14] However, because lives are finite and because present consumption is usually preferred to future consumption, an agent's concern for reputation will not completely eliminate moral hazard.

It is important to understand that moral hazard is *not* the same as fraud. Most interesting cases of moral hazard do not involve illegal behavior. It is not illegal for shareholders to take on riskier projects than the bondholders would like. Nor is it illegal for a manager to invest in projects with faster paybacks than shareholders would like. Moral hazard may involve fraud, but it need not. It will almost always involve ethical considerations.

Agency and moral hazard issues, like asymmetric information, pervade much of this book. The chapters that make heaviest use of these ideas are Chapter 3 in which we discuss the role of banks and other financial intermediaries, Chapters 5 and 6 on spot lending issues, and Chapter 10 on deposit insurance.

12. See Jensen and Meckling (1976) and Mirrlees (1976). James Mirrlees, a British economist, was one of the pioneers in models of moral hazard in economics, and was awarded the Nobel Prize in Economics for his contributions.

13. Chan and Thakor (1987) and Boot, Thakor, and Udell (1991) show how moral hazard can be reduced by collateral. Stulz and Johnson (1985) examine the relationship between collateral and firm value.

14. See, for example, Diamond (1989), Holmstrom (1999), Hirshleifer and Thakor (1990), John and Nachman (1985), Song and Thakor (2006), and Thakor (2005).

The warranty offered here can be viewed as a *signal* of quality. A (perfectly revealing) signal is one that enables the uninformed to infer which the informed agent knew privately *a priori*. For a signal to be useful it must be *informative*, and this requires that the signaling mechanism be incentive compatible. In turn, incentive compatibility requires that the cost of signaling must be negatively correlated with quality,[4] Michael Spence too was awarded the Nobel Prize in Economics for his contribution to the economics of asymmetric information. That is, it must be less costly at the margin for a higher quality seller to emit a given signal. The higher cost of signaling serves to deter the lower quality sellers from mimicking their higher quality counterparts. In our context, you can see that a warranty of \$12.50 imposes an expected liability of \$1.25 on the q_1 quality seller, \$6.25 on the q_2 quality seller and \$12.50 on the seller of lemons.

Note too that *in equilibrium* (that is, when each seller maximizes expected profit) the chosen signal is *costless* for the seller emitting it. Although the q_1 quality seller promises to pay \$12.50, he has only a 0.1 probability of having to pay, and since he collects \$11.25 upon selling the car, his cash inflow net of the expected liability is \$10. This is exactly what he'd have gotten *without* issuing a warranty, if we were in a "first best" world in which the quality of each car was common knowledge. Likewise, the q_2 quality seller's net cash inflow is \$5. Signals are *costless* in equilibrium. The reason for this, as you may have guessed, is that the seller is (correctly) compensated by the buyer for issuing the warranty, that is, cars with better warranties sell at higher prices. Such signals are called *nondissipative*[5] because the cost of the signal is a *transfer payment* from one party to the other, and there is no loss in the aggregate.

We can also have *dissipative* signals. To see this, suppose that instead of paying cash, the seller promises to reimburse the cost of repairing a portion of the damage. The q_1 quality seller promises complete coverage, the q_2 quality seller offers to absorb half the cost of repair, and the lemons owners choose not to participate. For every dollar it costs the seller to fix the damage, its value in terms of improved car quality is \$0.80. You can now easily verify that there exists a signaling scheme similar to the one derived previously that ensures truthful signaling by each seller, assuming that the seller is willing to accept a net payoff (after dissipative signaling costs are deducted) that is less than the car's worth. The q_1 quality seller's net receipt is less than \$10 and the q_2 quality seller's less than \$5. Each absorbs a signaling cost for which it is not compensated, that is, there is a net loss due to signaling.[6] For example, dividends can be a dissipative signal of future cash flows if they are personally taxed at a higher rate than capital gains (as was the case prior to the 1986 Tax Reform Act) and if external financing involves (transactions) costs that are avoided by financing with retained earnings.[7] Later in this book we will see other examples of dissipative signaling.

The concept of asymmetric information underlies much of what we discuss in this book, so you should expect to encounter it in more than a few of the remaining chapters.

4. See Spence (1973, 1974). Michael Spence too was awarded the Nobel Prize in Economics for his contributions to the economics of asymmetric information.

5. See Bhattacharya (1980).

6. If the seller is unwilling to bear the dissipative cost of signaling and the buyer will not bear it either, then a signaling equilibrium will fail to exist.

7. See, for example, Bhattacharya (1979).

Agency and Moral Hazard

It has been observed that the key distinction between man and machine is *moral hazard*.[8] First introduced in the insurance literature, this term describes situations in which the incentives of principal (the employer or the owner of the property) and agent (the employee or the person renting/using the property) diverge. A rational economic agent can be expected to maximize his own expected utility,[9] and where his self-interest conflicts with the principal's, the principal will suffer. The principal must therefore design a contract that will achieve a congruence between her goals and the agent's.

Examples of moral hazard abound. Consider automobile insurance. If you have a car that you know is worth \$500 and your collision insurance will pay you \$1,000 if the car is completely destroyed, you may be tempted to let your car roll down the hill and collide with an immovable object. Now you may never dream of doing this, but your willingness to spend on the maintenance of brakes may be subtly affected by your insurance policy. In any case, insurance companies cannot afford to assume that ethical or reputational considerations dominate their customers' behaviors.[10] This is one reason why we observe deductibles in insurance contracts. Coinsurance clauses are designed to share the risks and thereby bring the insured's incentives into closer alignment with those of the insurer.[11]

Moral hazard is also common in financial contracting among claimants in a corporation. Suppose you manage a firm and your goal is to maximize shareholder wealth. If you have risky bonds outstanding, you will not always choose investments that maximize the total value of the firm. Rather, you may choose projects that maximize the value of equity at the expense of the bondholders. This can be illustrated with the following numerical example.

8. Ross (1974).

9. We will refer to agents in the masculine and principals in the feminine.

10. Reputation enters via the customer's concern regarding future insurance premiums.

11. An interesting illustration of moral hazard is provided by the following report in *The Wall Street Journal* (WSJ) of October 10, 1990, titled, "More Car Owners are Scheming to Cheat Insurance Companies as Economy Falters."

"When a popular Dallas-area swimming hole developed a mysterious oil slick two months ago, it didn't take police long to discover something fishy was going on.

Littering the bottom of the abandoned stone quarry were 20 late-model automobiles, including a mint-condition 1990 Chevrolet Blazer. All of them had been reported stolen, and insurance companies had already paid off the owners. But contrary to claims in reports filed with insurance companies, most of the cars had keys in the ignition. And none of the vehicles had been stripped of fancy stereos, wheels or other easy to get accessories.

The police conclusion: The cars weren't stolen at all but had been dumped by their owners in what investigators say is one of the biggest 'car dunking' insurance scams in Texas history.

Hard figures aren't available, but most experts say 10% to 15% of all claim dollars paid out on car insurance result from some form of fakery. According to the Insurance Information Institute, that works out to between \$5.4 billion and \$8.1 billion of the \$54 billion in claims paid last year."

—Michael Allen, Staff Reporter of the *WSJ*

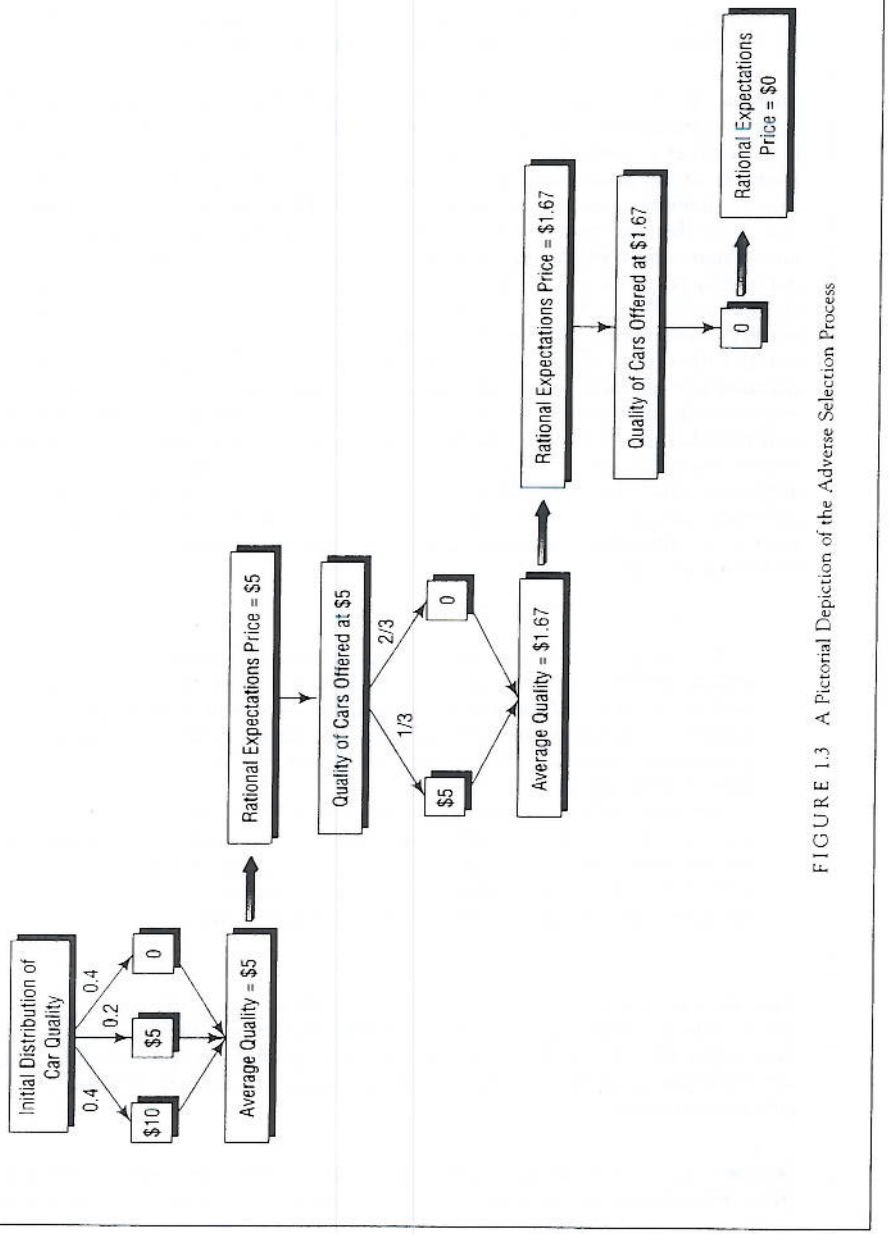


FIGURE 1.3 A Pictorial Depiction of the Adverse Selection Process

with warranties or guarantees. The seller of the q_1 quality car can announce that he will reimburse the buyer $\$W_1$ if his car fails, and the seller of q_2 quality car can announce that he will pay the buyer $\$W_1$ if his car fails. If buyers believe that *only* the owners of q_1 quality cars will promise a $\$W_2$ payment upon failure and that only the owners of q_2 quality cars will promise a $\$W_2$ payment upon failure, then they will make the appropriate inference and should be willing to pay prices that accurately reflect the qualities of the cars offered for sale. In order for such an indirect transfer of information to be effective, no seller should wish to mimic the strategy of a seller of a different quality car. Otherwise, buyers will eventually learn of the potential mimicry and the credibility of the signal will be destroyed.

Since the failure probability for a q_1 quality car is 0.1, the buyer should be willing to pay \$10 (the intrinsic worth of a q_1 quality car) plus 0.1 times W_1, the latter being the amount he expects to collect from the seller. Thus, the equilibrium price (P_1) of a q_1 quality car should be $\$10 + 0.1W_1$. Similarly, if the owner of a q_2 quality car follows his equilibrium strategy, the equilibrium price (P_2) of a q_2 quality car should be $\$5 + 0.5W_2$. To ensure that the q_2 quality car owner will not misrepresent himself as a q_1 quality car owner, W_1 should be set to satisfy

$$10 + 0.1W_1 - 0.5W_1 \leq 5 + 0.5W_2 - 0.5W_2. \qquad [1.15]$$

The left-hand side (LHS) of (1.15) is the expected payoff to a q_2 quality car owner misrepresenting himself as a q_1 quality car owner; he receives a price P_1 and has an expected outflow of 0.5 W_1 to pay the liability under the warranty. The right-hand side (RHS) of (1.15) is what the q_2 quality car owner gets if he follows his nonmimic strategy; he receives a price of P_2 and has an expected cash outflow of $0.5W_2$. When someone is indifferent between telling the truth and lying, it is conventionally assumed that truth-telling will be chosen. Thus, (1.15), which is referred to as an *incentive compatibility (IC) condition*, can be treated as an equality and we can solve it to obtain $W_1 = 12.5$. Incentive compatibility here means that the seller's incentives to maximize personal profit should be compatible with truthful representation of the car's quality.

The IC condition that ensures that the seller of lemons does not mimic the seller of q_2 quality cars can be similarly expressed as follows

$$5 + 0.5W_2 - W_2 \leq 0. \qquad [1.16]$$

Solving (1.16) as an equality yields $W_2 = 10$. It is straightforward to verify that the seller of q_2 quality cars will not mimic the seller of lemons under the described conditions, that is, q_2 quality cars will be offered for sale.

You can easily verify that this scheme guarantees that the seller of lemons will not mimic the seller of q_1 quality cars and that the seller of q_1 quality cars will not mimic either the seller of q_2 quality cars or the seller of lemons.

To summarize, we have produced a simple scheme of "warranties" that prevents market failure. The seller of q_1 quality cars promises to pay the buyer \$12.5 if his car fails; this enables him to sell his car for $10 + 0.1(12.5) = \$11.25$. The seller of q_2 quality cars promises to pay the buyer \$10 if his car fails; this enables him to sell his car for $5 + 0.5(10) = \$10$. The lemons are withdrawn from the market.

introduce. Suppose that in Example 1.2, we are given the prices of securities R_1 and R_2; recall that the price of each security is \$40. Moreover, R_1 pays off \$100 in state H and 0 in state L, whereas R_2 pays off 0 in state H and \$100 in state L. Let Pi_H and Pi_L be the prices of the Arrow-Debreu securities in states H and L, respectively. Then, the market price of security R_1 should be 100 times the price of the state H Arrow-Debreu claim, that is, $40 = 100\ Pi_H$, $Pi_H = 0.4$. Similarly, the market price of security R_2 should be 100 times the price of the state L Arrow-Debreu claim, that is, $Pi_L = 0.4$. We are now ready to price any security in this two-state economy. For example, the riskless bond in example 1.2, which pays \$50 in each state, should be priced at $50Pi_H + 50Pi_L = \$40$. A security that pays \$1,000 in state H and \$56 in state L should sell at $1000Pi_H + 56Pi_L = \$422.40$, and so on.

The concept of market incompleteness is used in Chapter 12 in connection with our discussion of financial innovation. Other applications can be found in chapters on off-balance sheet activities, securitization, and deposit insurance.

Asymmetric Information and Signaling

Economic transactions often involve people with different information. For example, the borrower usually knows more about its own investment opportunities than the lender does. Corporate insiders normally know more about the values of assets owned by their firms than shareholders. A doctor can be expected to be better informed about his or her own medical expertise than a patient.

The better informed economic agents have a natural incentive to exploit their informational advantage. Insider trading scandals on Wall Street illustrate how those with access to privileged information can profit, despite laws aimed at preventing such activity. Of course, those who are uninformed should anticipate their informational handicap and behave accordingly. It is this interaction between the inclination of the informed to strategically manipulate and the anticipation of such manipulation by the uninformed that results in distortions away from the "first best" (the economic outcome in a setting in which all are equally well-informed).

Problems of asymmetric information were brought to the forefront when George Akerlof (1970), who later went on to win the Nobel Prize in Economics for his contribution, sought to explain why used cars sell at such large discounts relative to the prices of new cars. The following example takes some shortcuts, but conveys the intuition of Akerlof's analysis.

Example 1.5 Consider a used car market in which differences in the care with which owners use their cars lead to quality differences among cars that started out identical. It is natural to suppose that the owner of the used car knows more about its quality than potential buyers. As an example, assume that there are three possible quality levels that the used car in question can have, $q_1 > q_2 > q_3 = 0$. If the quality level is q_3, the car is a lemon. Such a car would be priced as being worthless if buyers could correctly assess its quality. If the quality is q_2, the car has a value of \$5, and if the quality is q_1, the car is worth \$10. Assume that all agents are risk neutral and a buyer does not want to pay more for a car than its expected worth. In like vein, the car owner does not wish to sell at less than what the car is worth. Suppose that each car owner knows his car's quality, but buyers only know that cars for sale can be of quality q_1, q_2, or q_3. Faced with a given car, they cannot identify its precise quality. However, they believe that there is a probability 0.4 that the quality is q_1, a probability 0.2 that it is q_2, and a probability 0.4 that it is q_3. What will happen in such a market?

Solution If all cars are offered for sale, risk neutral buyers will compute the *expected* value of a (randomly chosen) car as $(0.4) \times \$10 + (0.2) \times \$5 + (0.4) \times 0 = \$5$. Hence, if the market is competitive, we would expect \$5 to be the market clearing price. However, at this price those who own cars with quality q_1 will refuse to sell. Thus, only cars of qualities q_2 and q_3 will be offered at \$5. However, buyers will anticipate this and revise their beliefs about the quality dispersion of cars in the market. They will now assume that if the selling price is \$5, the probability is $0.2/(0.2 + 0.4) = 1/3$ that the quality is q, and it is 2/3 that it is q_3. Thus, the expected value of a car drops to $(1/3)(5) + (2/3)(0) = \$1.67$. No cars will, therefore, be bought at \$5 (it cannot be a market clearing price). Now if \$1.67 is the price, those with cars of quality q_2 will drop out and the only cars offered for sale will be lemons. This process is called *adverse selection* and it results in the market clearing price being driven to zero. In other words, the demand for cars at any positive price is zero, and the market breaks down, as depicted in Figure 1.3. You should note a key assumption made in this example. All market participants have *rational expectations*. That is, uninformed buyers rationally anticipate what informed sellers will do at any given price and informed sellers rationally anticipate the demand buyers will have at that price. Hence, we don't need to go through a sequential process of price convergence to zero. No cars will be bought or sold.

The insight that asymmetric information can cause market failure was novel and striking. Its profound implications were quickly recognized to extend well beyond the used car market. Informational asymmetries were seen as being capable of causing markets to break down and thus possibly justify regulatory intervention by the government. Indeed, in the chapters that follow, we will examine banking regulation from this informational perspective.

However, calls for regulation based on Akerlof's analysis were too hasty. Market participants have the capability and incentives to deploy mechanisms to prevent market failure, and in any case market failure is the most extreme form of distortion created by asymmetric information. To see this in the context of our used car example, consider the following extension of that example.

Example 1.6 Suppose that cars of different qualities have different probabilities of engine failure within a given time period, and that these differences are reflected in their values of 0, \$5 and \$10. Suppose the failure probability is 0.1 for the q_1 quality car, 0.5 for the q_2 quality car, and 1 for the q_3 quality car. Do warranties have a role to play in this market?

Solution To prevent market failure, the sellers of better cars must somehow distinguish themselves from the sellers of lower quality cars. One way to do this would be

(*Continued on page 27*)

CHAPTER • 3

The What, How, and Why of Financial Intermediaries

"All essential knowledge relates to existence, or only such knowledge as has an essential relationship to existence is essential knowledge."

Soren Kierkegaard: *Concluding Unscientific Postscript*

Glossary of Terms

Securitization: The act of converting an untraded (debt) claim, such as a bank loan, into a traded security by issuing claims against it and selling these claims to capital market investors. Essentially, securitization is a form of direct capital market financing with the bank acting as an originator and repackager of the loan.

Fractional Reserve Banking: A banking system in which banks must hold a specified fraction of their deposit liabilities as liquid assets.

Fiat Money: A form of money, the acceptance of which is mandated by law.

The Market Model: A model that states that the return on a security can be partitioned into a fixed component (called "alpha"), plus a component which is a multiple (called "beta") of the return on the "market" portfolio, plus a mean-zero residual term.

DIDMCA: The Depository Institutions Deregulation and Monetary Control Act passed in 1980. See Chapters 11 and 12 for details.

The Law of Large Numbers: Roughly speaking, a principle that says that if we have an infinitely large number of random variables in a sample, all of which are drawn from the same probability distribution, then the average realized value of

the random variables in the sample will equal the statistical mean of the probability distribution from which they are drawn. Thus, if an individual divides his finite wealth equally across an infinitely large number of investments whose random payoffs are independent of each other, but are drawn from the same probability distribution, this individual's payoff from his investments will become (almost) *certain* and equal to the statistical mean of the probability distribution from which investment payoffs are drawn. A risk-averse individual would prefer to do this because it eliminates risk.

Event Study Methodology: A statistical approach commonly used in finance to evaluate the price impact of an event. The idea is to start with the assumption that the return on a stock can be described by the market model. Then, the next step is to estimate the values of alpha and beta by regressing the return on the stock against the return on the market for a sufficiently long time period prior to the event date and outside a 2- or 3-day time window around the event date. Given these estimated values, one can compute the average value of the residuals during the time window around the event date. If no new information was conveyed by the event, the average value of the residuals should be zero. If it is positive (negative), the event is interpreted as conveying good (bad) news.

Natural Monopoly: In some industries, due to economies of scale, the most economically efficient industry structure is to have only a single firm that is a natural monopoly.

Capital Requirements: The requirements that the bank keep a minimum amount of capital, consisting of equity, long-term debt, and other claims subordinated to deposits. See Chapters 2 and 11.

Portfolio Restrictions: Restrictions on the assets that banks can hold in their portfolios. See Chapters 2 and 11.

Introduction

As the following exchange between Levin and Sviyazhsky from Part III, Chapter 27 of Tolstoy's *Anna Karenina* indicates, most people know what banks and other financial intermediaries do.

> "Then what's your opinion? How should a farm be managed nowadays?"
> "What we have to do is to raise the standard of farming even higher."
> "Yes, if you can afford it! It's all very well for you, but... I'm not going to be able to buy any Percherons."
> "That's what banks are for."

As perceptive as this notion of banking is, we will need a deeper understanding of banks and other financial intermediaries in order to set the stage for the remaining chapters in this book. The simple view that banks exist to provide borrowing and lending services leaves us without answers to questions such as the following: (i) Why do we need *banks* to intermediate between borrowers and lenders, that is, why don't individual borrowers and lenders transact *directly* and avoid the cost of going through banks?[1]

1. A partial answer to this question was provided in Chapter 2.

(ii) What, if any, are the economies of scale in the production of financial services provided by banks, or, how large should banks be? (iii) Why do we regulate banks and other depository institutions so intrusively? (iv) If banks need to be regulated, *how* should they be regulated? (v) How should borrowers choose whether they should borrow from banks, or venture capitalists, or directly from the capital market?

To answer these and other questions, we need a framework that builds upon that provided in the previous chapter and illuminates the *essential* functions served by financial intermediaries. While we will not provide complete answers in this chapter to all of the questions posed above, our purpose is to provide a systematic way to think about these issues, so that we have a foundation for the discussions in subsequent chapters. The plan for this chapter is as follows. We begin with an anecdotal discussion of how a fractional reserve banking system arises from a simple goldsmith economy. After this informal discussion we provide a model of a bank that formalizes the goldsmith anecdote and helps us to understand the role of banks as well as the need to regulate them. These two sections provide answers to questions (i) and (iii) above, and a partial answer to question (iv). The next section introduces the fixed coefficient model as an extension of the goldsmith anecdote and examines its implications for monetary policy. The issue of economies of scale in the production of financial intermediation services is then taken up. This provides an answer to question (ii) above. Following this, we proceed to explain how banks can make nonbank contracting more efficient, and then we review empirical evidence in support of the view that banks are special. The ownership structure of depository institutions is analyzed next. We conclude with an examination of a borrower's choice of financing source to answer question (v) above.

Fractional Reserve Banking and the Goldsmith Anecdote

Fractional Reserve Banking

Chapter 2 explains what financial intermediaries do. We will now continue this discussion by examining how a rudimentary bank can evolve from a goldsmith, and how this leads to a theory of fractional reserve banking. What emerges too is a theory of bank regulation. According to this theory, regulation is an almost inevitable outgrowth of fractional reserve banking.

Modern banks produce *fiat* money on the basis of *fractional reserves*. These two facts account for much of the romance, mystique, and confusion surrounding finance. Laymen have difficulty understanding that money has value solely because of its universal acceptance as money.[2]

The fractional reserve aspect of banking is similarly vexing in that it seemingly involves sleight of hand. Fractional reserve banks fund themselves with liabilities that are convertible into cash on demand, but they hold only a fraction of such liabilities in the form of cash assets. Thus there is always some probability that withdrawals will exceed the available cash.

2. The acceptance of money is ultimately a social convention supported by the legal system, which recognizes money as an instrument for the legal discharge of debts. This view of money serves as the basis for arguing that seigniorage rightfully belongs to the community at large and should not be appropriable by private interests.

The evolution of monetary systems from commodity money—gold, silver, or whatever—to more abstract forms of money parallels the evolution of banking systems from warehouses, or 100 percent reserve banks, to modern fractional reserve banks. Both follow naturally from a collective desire to use scarce resources efficiently. However, these developments have side effects as well. The substitution of fiat for commodity money concentrates enormous economic power, for good or ill, in the hands of the monetary authority. Likewise, fractional reserve banking places enormous power in the hands of individual bankers, power to jeopardize the stability of the banking system in the pursuit of personal gain.

In what follows we shall explain the evolution of fractional reserve banking from its historical roots in warehousing. The explanation is stylized and anecdotal, and is meant to stress the natural aspects of the evolutionary process as well as the essential vulnerability of fractional reserve banking systems.

The Evolution of the Primitive Goldsmith Into a Bank

Think of a primitive setting in which gold is used as money—means of payment, or medium of exchange. By social convention, all debts are paid with gold and all purchases are made with gold. The system works well enough, but holding and transporting gold can be awkward. There is both a security problem and a convenience problem. The market response is to provide a warehousing service for gold. Hence the emergence of the goldsmith.

For a fee, the goldsmith provided secure storage facilities for gold. The owner of the gold would receive a warehouse receipt in exchange for her gold, with the understanding that the owner could present the receipt at her convenience to redeem the gold from the goldsmith.[3] The goldsmith's was a simple business. Like the furniture warehouse, the goldsmith provided safekeeping service for a fee. Simplicity itself!

Owners of gold gradually developed confidence in the goldsmith and gold flowed in and out of the goldsmith's coffers with tedious and profitable regularity. Whenever a gold owner wanted to make a purchase, she would travel to the goldsmith, withdraw the necessary gold and take it to the market. At the market, the gold would be exchanged for the desired goods and just as routinely, the seller of the goods would return the newly acquired gold to the goldsmith in exchange for a warehouse receipt.

As these trading and payment practices became more and more pervasive, and as the goldsmith's reliability became more and more established, repeated trips to the goldsmith were recognized as wasteful. Each time a purchase was desired, the buyer would need to run to the goldsmith for gold, only to have this trip repeated by the seller, who would return the gold from whence it came.

Ultimately, the warehouse receipt passes from the buyer to the seller, and the only purpose served by the two trips is to test the goldsmith's integrity. But as the goldsmith's reputation for integrity grows with time and experience, the need for these trips seems increasingly unnecessary. Gradually, trade is effected with the exchange of warehouse receipts and the gold remains undisturbed in the goldsmith's vault. But the willingness to accept warehouse receipts in lieu of gold rests on the belief that the gold is available on demand. Any suspicion of the goldsmith will undermine the use of

3. When transferable, it is the ownership of the receipt that governs the redemption.

the receipts as means of payment. But so long as the goldsmith can project confidence, there is a saving to be had by avoiding the trips to and from the goldsmith.

Seen from the vantage point of the goldsmith, the growing use of receipts as means of payment means smaller flows of gold into and out of the coffers. One can imagine a time series of data points that describe the gold holdings of the smithy through time. As the use of receipts gradually replaces gold, the goldsmith's gold inventory becomes less and less volatile. In the limit, as the receipts totally displace the gold, the goldsmith's inventory remains practically unchanged through time, unless newly mined gold flows into the system, or other extraordinary occurrences take place. It gradually dawns on the goldsmith that it is not really necessary to have a unit of gold for each outstanding receipt. This idea must have come as a revelation, an epiphany. To be sure, the strait-laced would recoil at the idea of issuing more receipts than one had gold, but if no one ever withdraws the gold, then what possible harm?[4] The naughty possibility of printing extra warehouse receipts changed the world. This discovery was the banking equivalent of the Newtonian Revolution, every bit as important to banking as gravity was to physics.

Instability of the Fractional Reserve Bank

The extra receipts could not be distinguished from their more authentic counterparts and they consequently served as means of payment as readily as did the authentic (those whose issue was occasioned by a deposit of gold) receipts. The extra receipts were loaned to borrowers and earned interest. Assume that these loans are illiquid, that is, they cannot be redeemed on demand, but rather must be held to maturity in order to realize their full value. This means that the goldsmith is providing a key *liquidity transformation* service by issuing liquid claims to depositors that are backed by illiquid loans to merchants. The pedestrian goldsmith was thus transformed from a warehouse clerk into a banker! To see this, consider the following before-and-after balance sheets.

Goldsmith (Before)		Goldsmith (After)	
Gold 100 oz.	Receipts 100 oz.	Gold 100 oz. Loan 10 oz.	Receipts 110 oz.

Notice that after the goldsmith crosses the Rubicon (becomes a banker), his liabilities of 110 ounces exceed his capability to satisfy them in the unlikely event that all receipt owners should seek to convert to gold simultaneously. This potential failure is because loans are *illiquid*.

Therefore, inherent in the lending is a potential catastrophe—insolvency of the goldsmith. Of course, if the receipt owners almost never withdraw their gold, the probability of insolvency is small, perhaps very small. However, and this is critical, the risk of ruin is endogenous. That is to say, the goldsmith chooses the probability of insolvency with his choice of how many extra receipts to print, or equivalently, with his choice of how many loans to make. Each extra receipt printed and loaned earns interest and so the temptation to print receipts is limited only by the goldsmith's concern for remaining solvent. He walks the knife-edge between avarice and anxiety.

4. In a rational expectations equilibrium, the gold owners would anticipate this behavior of the goldsmith and adapt (redeem randomly and sufficiently frequently) to avoid being exploited by the goldsmith. But for present purposes, let us ignore this.

Each extra receipt increases income, but at the same time increases the probability of insolvency; insolvency, of course, destroys the goldsmith's reputation and with it his ability to circulate and lend warehouse receipts.

Thus we see how the discovery of fractional reserve banking was a rite of passage, a loss of innocence. Notice, however, that conditional on the loans being repaid, the goldsmith holds assets equal to the value of his liabilities. Thus what we have here is a liquidity issue. The goldsmith can and will pay off all receipt holders, given adequate time and good loans. Nevertheless, the promise is to *pay on demand*, and this most assuredly cannot be done in *all* states of nature.

This is the essence of fractional reserve banking and its essential vulnerability. Such a system evolves quite naturally given maximizing behavior on the part of rational economic agents.

Regulation as a Stabilizing Influence

Left to its own devices, this kind of banking system is subject to periodic collapse. However, experience with fractional reserve banking eventually led to the discovery of a rather simple and straightforward remedy. Since the Achilles heel of the system is the illiquidity of the loans, bank runs could be averted if these assets could be liquefied. What was needed was a *bank for goldsmiths* that could lend against the collateral of a goldsmith's loans during those infrequent occasions of extraordinary redemptions. Indeed, in the 19th century this was achieved in the U.S. through commercial bank clearing houses (CBCHs), which were private arrangements between banks that agreed to put their *combined* resources (the CBCH) behind each member in times of unanticipated liquidity drains. (See Chapter 9 for more on CBCHs.) Of course, such a bankers' bank would need virtually unlimited capacity, together with a commitment to the continuity of the system. The private arrangements did not possess such unlimited capacity, and this provided the rationale for a central bank to serve as a lender of last resort to the community of bankers. Since the central bank, which was typically government-owned, had the privilege of printing (or otherwise creating) money, the issue of limited capacity evaporated.

One more point deserves emphasis in connection with the evolution of a fractional reserve banking system with a central-bank-based lender-of-last-resort facility. Absent the central bank, there will always be a self-imposed limit on the volume of extra receipts printed. The fear of failure, loss of reputation, and the consequent inability to continue to lend warehouse receipts will discipline the inclination to expand lending indefinitely. Whatever this self-imposed limit, however, the introduction of the central bank acting as a lender of last resort will weaken the goldsmith's restraint. If the goldsmith knows that he can borrow against his otherwise illiquid loans, he will make more loans than if he could not use the loans as collateral. This is clear and obvious; and it is true even if the central bank charges a very high rate of interest for such emergency borrowings. Note that the interest rate for such loans is infinite in the absence of the central bank. Thus, the central bank introduces a kind of moral hazard, and this moral hazard is typically addressed by imposing cash asset reserve requirements that effectively limit the volume of a bank's lending on the basis of its cash assets. This is perhaps the most basic of prudential regulation. The point is that regulation is endogenous. It is responsive to a moral hazard arising from the introduction of the central bank as a lender-of-last-resort, which in turn is a response to a vulnerability inherent in fractional reserve banking. In turn, fractional

reserve banking is a natural response to the transport costs and security concerns in a *laissez-faire* world of commodity money.

A Model of Banks and Regulation

That the very nature of banking necessitates regulation can also be seen in the perspective of a model in which money—rather than gold—is used as a medium of exchange. We will now develop in the box below a model that formalizes the anecdotal development of the previous section and also highlights some of the underlying informational assumptions in the analysis. The intuition is very similar to that in the earlier section.

The two-period model developed below is very simple.[1] It makes some assumptions that are not rigorously justified. Our intent is to give a broad-brush, intuitive treatment of how banks arise even in primitive economies and why it is necessary to regulate them. Before developing the model, we provide a summary of the notation used in Table 3.1

TABLE 3.1 The Notation

Notation	What it Means
y	Depositor's income in each period.
c	Depositor's consumption from income in each period.
s	Amount deposited in each period.
ϕ	Fee charged to depositor for safekeeping of deposits.
$\hat{\phi}$	Personal cost of safeguarding deposits.
α	Fraction of deposits withdrawn.
n	Number of depositors.
m	Number of merchants.
K	Merchant's cash flow.
K^*	High value of merchant's cash flow when K is random.
M	Loan to merchant.
p	Probability of theft.
r	Rate on return on bank's loan to a merchant.
b	Bank's cost of monitoring merchants.
u	Probability that $K = 0$, as assessed at date 0.
u_1	Value of u, as per updating at date 1.
u_h	High value of u_1.
U_ℓ	Low value of u_1.
L	Liquidation value of merchant's investment.
f	Amount depositors must spend to ensure that the bank safeguards and monitors.
j	Number of banks.

(Continued)

The Model: Consider an economy in which individuals are unsure of how safe their personal wealth is from theft. Thus, it pays to safeguard it. The individual can either safeguard it himself or he can pay someone else to do it. It is easy to imagine that not everybody is equally skilled in the art of safeguarding. So if you believe others are more skilled in safeguarding, you may wish to entrust safeguarding of your wealth to someone else, even though this involves paying a fee.[2] Since we will eventually reach this conclusion anyway, let us refer to you (the person who wishes to have his personal wealth safeguarded) as the depositor and the entity that safeguards your wealth as the bank. For now let us suppose there is only one depositor (n = 1) and only one bank (j = 1).

The depositor has an income of $y in each period, of which $c goes to personal consumption and $(y − c) = $s goes to savings. These savings must be safeguarded. For now suppose there is nothing that the bank can do with this money except safeguard it. Let $ $\phi > 0$ be the fee that the bank charges to safeguard the depositor's savings. Safeguarding by the bank guarantees that the wealth will not be stolen. Also suppose that the depositor wishes to have his wealth safeguarded for only one period. Assuming that the discount rate is zero for everybody,[3] we see that the depositor's consumption at the start of the next period will be $s - \phi$ (his net saving in the first period) plus $y (his income in the second period). Since the depositor is paid $s − ϕ for depositing $s, the interest rate on his deposit is

$$(s - \phi - s)/s = -\phi/s < \qquad .1]$$

If a negative interest rate surprises you, remember that our bank cannot make any loans and is providing the depositor a costly service. Assume for now that the bank must keep 100 percent reserves against deposits and that the depositor will fully withdraw at the end of the first period.

The Desirability of a 100 Percent-Reserves Bank: Suppose the probability of theft is p and it would cost the depositor $\hat{\phi} > \phi$ to safeguard his wealth to the extent that the probability of theft is eliminated. Thus, a necessary and sufficient condition for personal safeguarding to be optimal is that

$$\begin{aligned} &s - \hat{\phi} > (1 - p)s \\ &\text{or } \hat{\phi} < ps \\ &\text{where } 0 < p < 1. \end{aligned} \qquad [3.2]$$

We will assume that (3.2) is satisfied. Clearly, since $\phi < \hat{\phi}$, the depositor will prefer to have the bank safeguard his wealth.

Note that in stipulating that the bank charges the depositor exactly what it costs the bank to safeguard, we have assumed that there is perfect competition[4] between banks that can all safeguard s at ϕ. Suppose now that n > 1, so that there are possibly many depositors. It would be natural to assume that there are economies of scale in safeguarding, that is, it should cost less per dollar to safeguard $ns as opposed to safeguarding $s. For example, one armed guard may be able to safeguard $100,000 just as easily as he can safeguard $1,000. Indeed, if we were to assume that the cost of safeguarding $ns is less than n\phi$, the case for a large bank would be compelling, and we could even assume than $\hat{\phi} = \phi$, that is, no single individual is any more skilled than

another in protecting wealth. But we will assume that there are no scale economies in safeguarding. In a sense, this makes our task harder, but it helps to reduce notation.

Suppose first that all the n depositors will surely withdraw at the end of the first period. In this case, it is easy to see that the interest rate will still be $-\phi/s$. A more interesting and natural case, however, is one in which *not* all depositors will withdraw at the end of the first period. Suppose a fraction α (where $0 < \alpha < 1$) of depositors will withdraw at t = 1 (the end of the first period) and the remaining fraction $1 - \alpha$ will withdraw at t = 2 (the end of the second period). For simplicity, we assume that α is known with certainty.[5] The sequence of events is described in Figure 3.1 below.

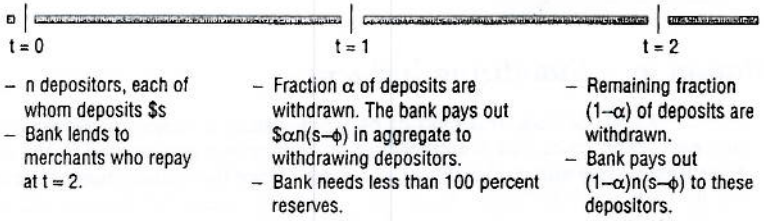


FIGURE 3.1 Sequence of Events

A Bank That Borrows and Lends: If the bank cannot invest any of the deposits it receives, then funds will lie idly in the bank. Note, though, that there is an opportunity to invest in this case.[6] At t = 1, the bank only needs to have $\alpha n(s - \phi)$ to meet deposit withdrawals. Suppose now that it is possible for the bank to make investments at t = 0, but that these investments will pay off only at t = 2. Let r be the rate of return to the bank on these investments.[7] We can imagine that the investments are loans to merchants who want to finance the setting up of shops, but do not have any funds of their own. Each merchant needs $M, where $M > s$, so that if the merchant were to borrow directly from depositors, he would need to approach more than one depositor (in fact, he would need to approach *M/s* depositors). Further, there is a *moral hazard* problem in dealing with the merchant in that he has a preference for absconding with the $*M* he borrows rather than setting up a shop. If his actions are not monitored, he will abscond and the lender will not be paid back at all. However, at a cost of $b it is possible to monitor the merchant so that he indeed puts his borrowed funds to the stated use of setting up a shop that will generate some cash flow of $K > $M(1 + r)$ at t = 2. As a start, let us suppose that $K is a sure cash flow. We will introduce uncertainty shortly.

First consider the merchant's problem if he approaches M/s depositors directly. His net expected payoff will be

$$K - M(1 + r) - (b \bullet M/s \qquad .3]$$

since, in addition to interest, he will be charged for monitoring. Each depositor will have to individually monitor the merchant since none can rely on his cohorts to do so.[8] Now, if the merchant approaches a bank, which in turn acquires $M in deposits from *M/s* depositors, we will have a different outcome. The bank's monitoring cost

(Continued)

will be \$b. If the bank charges the merchant exactly what it costs the bank to monitor, then the expected payoff to the merchant will be

$$K - M(1 + r) - b. \qquad [3.4]$$

Comparing (3.3) and (3.4) we see that the merchant is clearly better off going to the bank.

Since the merchant pays the bank only at $t = 2$, the bank will have to make sure that it will have enough money at $t = 1$ to pay off depositors who withdraw then. Suppose there are m merchants (borrowers) and n depositors. Then, the bank loans out \$mM and takes in \$ns in deposits. Let $ns > mM$ (this will be shown to be necessary in a moment). Since \$mM are loaned out, the bank doesn't need to worry about safeguarding that money from outright theft (it just needs to monitor the merchants it lends to). Thus, \$(ns − mM) must be safeguarded. The safeguarding cost is $(ns - mM)\phi/s$, since it costs ϕ/s to safeguard \$1.

Hence, the bank promises to pay depositors

$$ns - (ns - mM)\phi/s \qquad .5]$$

in the aggregate if it does not pass along to the depositors any of its profits from lending to merchants. Since a fraction α of deposits are withdrawn at $t = 1$, those depositors get $\alpha[ns - (ns - mM)\phi/s]$, which you will notice is more (by an amount $(mM\phi/s)$ than what these depositors received previously. That is, the fact that part of the money is being loaned out instead of being kept in the bank's vault itself economizes on safeguarding costs. Although the loaned money must be monitored, these monitoring costs are paid by borrowers, so that depositors realize a saving in safeguarding costs.

To ensure that the bank will have sufficient funds to meet deposit withdrawals at $t = 1$, it must choose m to satisfy

$$\alpha\,[ns - (ns - mM)\phi/s] = ns - mM - (ns - mM)\phi/s - mb. \qquad [3.6]$$

To understand (3.6), note that the left-hand side is the amount the bank must pay out to those depositors who withdraw funds at $t = 1$. On the right-hand side, ns−mM is the amount of money the bank has left over in reserves after it is through lending to the m merchants. From this it must spend an amount (ns−mM) ϕ/s to safeguard its reserves and an amount mb to monitor the m merchants.[9] Solving (3.6), we get

$$m = (1 - \alpha)ns(s - \phi)/\{M[s - (1 - \alpha] + bs\} \qquad [3.7]$$

Thus, as long as the bank lends to exactly as many borrowers as stipulated in (3.7), there will be no risk of withdrawals exceeding the bank's available cash reserves at $t = 1$.

Note now that the bank makes an aggregate net profit of mMr on its lending activities. This is because it is being compensated exactly for its monitoring cost by borrowers, and its safeguarding cost by deposit interest rate, although higher than $-\phi/s$ (as in the previous case when all deposits were idle), is still negative. This

positive profit will attract entry by competing banks, and the resulting competition for depositors' funds will drive up the deposit interest rate. In a competitive equilibrium, each bank will earn zero profit. This will happen when the bank's profit of mMr is divided equally among the n depositors, so that each depositor gets

$$\frac{ns - [(ns - mM)\phi/s] + mMr}{ns}$$

per dollar of deposits. Thus, the deposit interest rate is now

$$\frac{ns - [(ns - mM)\phi/s] + mMr}{ns} - 1$$
$$= \frac{mMr - [(ns - mM)\phi/s]}{ns} \qquad [3.8]$$

If we assume that r is high enough to ensure that the numerator in (3.8) is positive, then the depositors get a positive rate of interest on their deposits.

We have taken you through a sequence of steps to show how a bank, like the goldsmith in the previous section, can develop from a simple caretaker of other people's wealth into an institution that borrows and lends money. As you must have noted, informational problems play a key role in bringing our bank to life. Banks solve two types of moral hazard problems in our simple world. First, they help to cope more efficiently with the "social" moral hazard problem of theft. Second, they also help to cope more efficiently with moral hazard in lending, which, as you know from Chapter 1, is a type of agency problem.

Do We Need to Regulate This Bank?: So far, however, there has been no need for a regulator. But that is simply because we have made numerous strong assumptions. One of them is that it is possible to monitor merchants so efficiently that they'll always repay their debts fully if they are monitored. In reality, merchants may sometimes have poor cash flows even if they do their best. That is, suppose that, viewed at $t = 0$, their cash flow K is a random variable that is 0 with probability u and K^* with probability $1 - u$. We'll assume that setting up a shop is a positive net present value (NPV) exercise for the merchant, so that

$$(1 - u)K^* > M(1 + r). \qquad [3.9]$$

Suppose that this in itself does not affect the behavior of depositors in terms of their withdrawal policies. But at $t = 1$, depositors may learn something more about the likelihood that merchants may fail. For simplicity, assume for now that merchants have perfectly correlated prospects, so that they all either fail ($K = 0$) or succeed ($K = K^*$). Let us refer to the updated probability of failure that depositors assess at $t = 1$ as u_l. If there is good news, $u_l < u$ (the probability of failure they assessed at $t = 0$) and if there is bad news, $u_l > u$. We can think of u as the expected value of u_l assessed by depositors at $t = 0$. Suppose u_l can take one of two values: $u_l = u_h$ for bad news and $u_l = u_\ell$ for good news, where $u_h > u_\ell$. Suppose that those depositors who intended to withdraw at $t = 2$ will in fact change their minds and withdraw at

(Continued)

$t = 1$ if they get bad news[10], that is, if $u_l = u_h$. If they get good news, they'll withdraw at $t = 2$.

The bank now faces a problem. If depositors get bad news, all depositors withdraw at $t = 1$. The bank will have insufficient funds to meet withdrawals (unless it keeps 100 percent reserves and does not lend to any merchants). Suppose that in this case the bank is empowered to call back all of its loans prematurely and this forces merchants to liquidate their businesses prematurely. Let L be the liquidation value of the merchant's shop at $t = 1$ (which, for simplicity, is independent of the information received by depositors at $t = 1$). Assume L is a very small number (much smaller than K^*). So, if all depositors wish to withdraw funds at $t = 1$, and if the bank proceeds to lend exactly the same amount at $t = 0$ as it did in the previous case, then there will only be $\$\alpha[ns - (ns - mM)\phi/s] + mL$ to pay depositors. Moreover, the premature liquidation of merchants' shops will be socially inefficient if L is so small that $L < (1 - u_h)K^*$. This is similar to the illiquidity problem of the goldsmith.

There is no way that the bank can prevent this unless it keeps all of its deposit funds idle, in which case it doesn't matter when depositors withdraw. However, this would *not* be fractional reserve banking; it would hardly be a bank as we know it. This is where a government regulator can help. Suppose it agrees to insure all deposits for the full promised payment by each bank. Then we see that those depositors who originally planned to withdraw at $t = 2$ have no reason to change their minds since the value of u_l is now irrelevant to them; the deposit insurer has made their claims risk free! That is, this form of regulation makes banking viable when it otherwise could not have been.

This seems to be a wonderful solution and it definitely has its merits. But lest we get carried away with its virtues, let's pause and complicate things a bit more. Since banks are competitive and earn zero profits, they may wish to underspend on either safeguarding or on monitoring borrowers. Once the terms of their loan and deposit contracts are set, they could profit from spending less on safeguarding and monitoring than originally promised. Depositors will rationally anticipate this moral hazard and try to prevent it. Suppose that each depositor could spend a small amount of money, say \$f, to make sure that the bank expends the promised resources on safeguarding and monitoring. We can show, given appropriate assumptions, that depositors will find it in their own best interest to do so.

1. This model has some features found in Millon (1983). Other papers dealing with the existence of financial intermediaries are Leland and Pyle (1977), Campbell and Kracaw (1980), Diamond (1984), Ramakrishnan and Thakor (1984), Millon and Thakor (1988), Boyd and Prescott (1986), and Allen (1990).

2. Naturally, this fee should be less than what it would cost you to safeguard your own wealth with the same efficacy.

3. This is a harmless assumption and can be easily dropped without affecting this analysis.

4. For those of you well-versed in different notions of competition in economies, we have in mind Bertrand competition here.

5. We will discuss later what happens if α is random.

6. Actually, even in the previous case in which all deposits are withdrawn at $t = 1$, the bank could invest at $t = 0$ in assets that pay off at $t = 1$.

7. We will not go into the details of how r is determined.

8. It is obvious that we cannot have an equilibrium in which no depositors monitor, because then it pays for at least one to monitor. To justify an individual depositor's decision to monitor, we must assume that there is some uncertainty that some depositors will not monitor (otherwise, every depositor will wish to "free ride" on the monitoring of his cohorts). One way to do this is to assume that *each* depositor believes that there is a random fraction θ of the remaining $(M/s) - 1$ depositors who are simply incapable of monitoring, but no one (except those incapable depositors themselves) can identify these depositors. Thus, each of the depositors will still charge for monitoring but will not spend \$b. Suppose θ can be 0 with probability q_0 and 1 with probability $1 - q_0$ (when $\theta = 1$, each depositor who can monitor believes that he is pivotal in that no one else will monitor). Then, if a depositor who monitors chooses not to do so, his expected payoff will be (he always assumes that all other depositors capable of monitoring will indeed monitor) $b + q_0 s(1 + r) - s = b - (1 - q_0)s + q_0 sr$. And if he chooses to monitor, his expected payoff will be $s(1 + r) - s = sr$. Thus, it is a (Nash) equilibrium to monitor if $sr > b - (1 - q_0)s + q_0 sr$ or if $b < (1 - q_0)s(1 + r)$. Thus, if the uncertainty about incapable depositors is sufficiently large in the mind of each capable depositor (that is, $1 - q_0$ is sufficiently high) and if the monitoring cost b is low relative to the payoff $s(1 + r)$ from successful monitoring, each capable depositor will monitor in a Nash equilibrium.

9. We are assuming here that safeguarding costs are paid just after $t = 0$ and monitoring costs are paid just before $t = 1$. Note that since the merchants repay the bank only at $t = 2$ and monitoring must proceed at $t = 1$, the bank must initially pay the necessary monitoring costs and then recover these costs from borrowers at $t = 2$ through a loan interest rate that is grossed up to reflect this cost.

10. Let us not worry about why they might wish to do this. We want to give you an idea of the underlying concepts without being too rigorous. It is possible to make these ideas work more rigorously.

Summary: Thus, one way to prevent bank runs and instability is for the government to provide deposit insurance, which is an alternative to the lender-of-last-resort (discount window) facility provided by the regulator in our earlier goldsmith example. But there is a fly in this ointment. When there is deposit insurance, why should any depositor care about whether the bank safeguards and monitors with the requisite vigilance? Each depositor's payoff is guaranteed and independent of the bank's actions. Hence, none will find it personally profitable to spend anything on watching over the bank to ensure that the bank expends the promised resources in safeguarding and monitoring the merchants it lends to. In other words, deposit insurance weakens or even destroys the private market discipline imposed on banks. The burden of keeping the bank in check shifts now from the market to the regulator. To achieve its objective, the regulator will have to come up with ways to dissuade the bank from exploiting the deposit insurance umbrella. In other words, the moral hazard engendered by one form of regulation, namely deposit insurance, creates the need for other forms of regulation (such as capital requirements, portfolio restrictions, and so on).

We have now completed the story we set out to tell in this section. Regulation is not just the outcome of some political agenda. It arises quite naturally from the very forces that give rise to banks. Once regulation arises to instill public confidence in banking and make banks viable entities, it creates its own moral hazards that necessitate further regulation.

The Macroeconomic Implications of Fractional Reserve Banking: The Fixed Coefficient Model

In this section we examine the implications of fractional reserve banking for monetary policy. The discussion developed here formalizes some of the macroeconomic implications of the goldsmith anecdote presented earlier.

The Fixed Coefficient Model

The Fixed Coefficient Model (FCM) is the standard textbook description of the banking firm and industry; it emphasizes the asset-transformation function of financial intermediaries. The bank's effort to maximize its profit is captured only implicitly. Consider a bank's balance sheet

Bank Balance Sheet

R	D
M	E

where R is the reserves of the bank comprised of deposits held at the central bank, M is the bank's earning assets (loans to merchants), D is the bank's deposit liability (think of this as n × s in the context of the model in the previous section), and E is the bank's equity. We can now write the balance sheet identity for the bank as:

$$R + M = D + E. \quad [3.10]$$

Moreover,

$$R = rD, \text{ with } 0 < r \leq 1. \quad [3.11]$$

Equation (3.11) represents the fact that banks hold cash or liquid asset reserves proportional to deposits in order to insure against deposit withdrawals and/or to satisfy legal reserve requirements. The fixed coefficient, r, can be interpreted either as a legal reserve requirement or a voluntary behavioral parameter (that is, reserves that the bank chooses to voluntarily hold). Actually, it should be interpreted as the greater of the two. In any case, the parameter relates to *liquidity or withdrawal risk*. That is, it is the bank's safeguard against a fraction (α in the context of the model in the previous section) of deposits being unexpectedly withdrawn. Next, we have

$$E = eL, \text{ with } 0 < e \leq 1. \quad [3.12]$$

Equation (3.12) represents the fact that banks hold capital reserves in some fixed proportion, e, to loans in order to protect against *insolvency or default risk*. The parameter e can be interpreted as a regulatory capital requirement and/or a voluntary behavioral parameter, or, more accurately, the greater of the two.

An Illustration of the FCM

Let us now consider the FCM in a (competitive) banking industry with zero equity (e = 0) where banks have only two assets (reserves held in the form of deposits at the Federal Reserve and loans to the public) and one liability (customer deposits). We shall further assume a 20 percent effective legal reserve requirement (r = 0.2). The assumption that e = 0 is an extreme representation of the assumption that the capital requirement is not binding.

Now suppose Bank A receives a $1,000 deposit.

Bank A

Required Reserves 200	1000 Deposits
Excess Reserves 800	
Total Reserves 1000	

Since it has excess reserves of $800 and since it earns nothing on either its required reserves or excess reserves, the bank seeks to eliminate its excess reserves by making a loan of $800:

Bank A

Total Reserves 1000	1000 Deposits
Loan 800	800 Deposit

The funds loaned by Bank A, although possibly initially deposited with Bank A, are soon withdrawn and deposited in another bank, say Bank B. This leaves Bank A with

Bank A

Required Reserves = Total Reserves 200	1000 Deposits
Loans 800	

But Bank B has

Bank B

Required Reserves 160	800 Deposits
Excess Reserves 640	

and Bank B now lends away its excess reserves, so that:

Bank C

Required Reserves 128	640 Deposits
Excess Reserves 512	

The $640 loaned by Bank B is now deposited in Bank C. The process continues *ad infinitum*. At the Federal Reserve, the initial deposit would be a credit of $1,000 to Bank A.

Federal Reserve	
	1000 Deposit A

What is the offsetting asset (liability) entry?

When the $800 is withdrawn from A and deposited in B, the Federal Reserve would show

Federal Reserve	
	200 A
	800 B

Notice that the original reserve creation (the $1,000 deposit received by Bank A) spurred deposit expansion, and the deposit expansion redistributes the reserves across the banking system. However, the deposit expansion does not affect the level of reserves in the banking system. In fact, deposit expansion absorbs reserves. What this illustration of the FCM shows is that the bank's incentive to hold reserves—either voluntarily to protect against unanticipated deposit withdrawals or to satisfy a regulatory reserve requirement necessitated by the moral hazard created by the lender-of-last-resort facility—results in less lending than would be possible without reserve requirements. Moreover, it also affects the redistribution of liquidity throughout the entire banking system. This has macroeconomic implications that we explore below.

The FCM and Monetary Policy

The FCM helps us to understand the basic elements of how monetary policy works. There are three major tools of monetary policy: (i) open market operations, (ii) reserve requirement changes, and (iii) discount rate changes. These three tools are used in varying degrees to influence the stock of money and interest rates.

Open market operations are sales and purchases of government securities (Treasuries) by a special committee of the Federal Reserve. These sales and purchases affect the amount of reserves available to banks and thus, as indicated in previous subsections, the amount of lending. To see this, suppose the Fed buys $1,000 in Treasury securities from the nonbank public. Then the nonbank public's balance sheet will be

Public	
Bonds − $1,000	Liabilities unchanged
Deposits of cash in Bank A + $1,000	

and Bank A's balance sheet will be

Bank A			
Required Reserves	200	1000 Deposits	
Excess Reserves	800		

The $800 is now available to Bank A for lending. This means that the initial open market operation of purchasing Treasuries leads to an increase in lending by banks. Another way to view this is that the government has reduced public debt (by buying back government securities) and facilitated an increase in private credit. The open market operation of selling government securities has the opposite effect.

It is obvious that a change in reserve requirements will also affect bank lending. Any increase in reserve requirements will reduce the amount of deposits available for lending, and any reduction in reserve requirements will increase the amount of deposits available for lending. Thus, when the Federal Reserve desires to implement a contractionary monetary policy (to cool down inflation, for example), it can raise reserve requirements; similarly it can lower reserve requirements when it wishes to stimulate the economy.

Finally, the discount rate, which is the rate charged by the Fed to member banks for short-term borrowings from the Federal Reserve, also affects monetary expansion/contraction. By raising the discount rate, the Fed makes it more costly for banks to borrow and build up reserves, and therefore effectively reduces the reserves available to banks. This reduces lending. Likewise, a lowering of the discount rate facilitates increased lending.

This analysis is predicated on the "classical" assumption that the binding constraint on bank lending is the reserve requirements. If the capital requirements e [recall equation (3.12)] were binding instead, the effects of monetary policy can be very different indeed, as we will see in Chapter 10.

Large Financial Intermediaries

The theories from which we borrowed some of the ideas in the previous section suggest that financial intermediaries should be very large. These arguments are based on diversification. They explain why banks should be large. Similar intuition applies to nondepository financial intermediaries as well. In this section we develop this argument. We focus on the basic intuition; the mathematics can be found in Appendix 3.1. It leads to a rationale for *nondepository* financial intermediaries like investment banks, Standard & Poor's Value Line, credit rating agencies, financial newspapers, Moody's check guarantee services, portfolio managers, econometric modelers, consultants, and accounting firms.

What the theoretical research has shown is that F.I.s are optimally infinitely large regardless of whether they are brokers or asset transformers. That is, an F.I. is a "natural monopoly." We explain why below.

Brokerage as a Natural Monopoly: Consider a broker that specializes as an information producer. One problem that the broker's customers must be concerned about is that of information reliability. This is a key issue in information production. How do these customers know that the information the broker provides is accurate and reliable? One possible way to determine this is for customers to noisily assess the

reliability of the information provided by the broker, and compensating the broker more when information is judged to be more reliable. This can be done either via reputational mechanisms – attaching higher reputation for reliability to a broker whose past information has turned out to be higher quality – or by comparing the broker's information to that available from other sources.

Now, if we are dealing with a single information producer, it can be quite costly to ensure that he will use reliable information, even if we can have a noisy assessment of this reliability. This becomes a little less costly if we are dealing with a producer who is a member of a *team* of information producers because then, by producing reliable information, *each* producer benefits not only himself (by making it more likely that he will obtain higher compensation) but also the team, and a share of the team's benefits accrues to each individual producer. This is an effective mechanism as long as the team members can monitor each other to ensure that nobody gets a "free ride." As the size of the team grows, more and more independent payoffs of individual producers are being pooled together before being divided equally among the team members, so that the resulting diversification reduces the risk in each member's compensation. The risk-averse information producers are thus made better off and they demand less compensation on an expected value basis to produce information. This makes the buyers of information better off. And the benefit keeps growing as the broker gets larger. That is, brokerage is a natural monopoly.

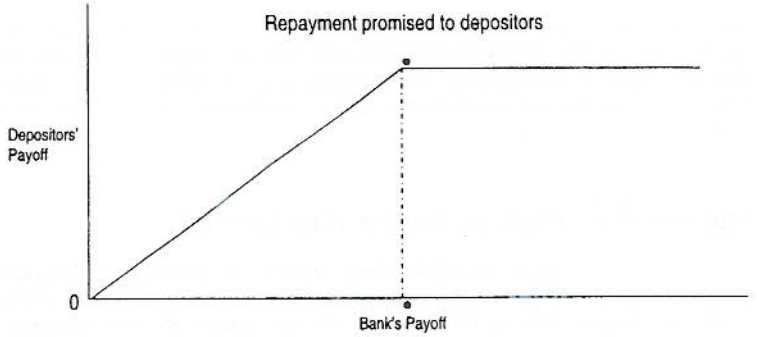
Another economic benefit from growing large comes from *information reusability*, which was discussed in Chapter 2. When information is cross-sectionally reusable, the larger the number of information producers in the intermediary, the greater is the benefit of information reusability. The reason is that information can be reused by a greater number of information producers within the intermediary, and yet the cost of acquiring information needs to be incurred only once.

A strong implication of this analysis is that investment banks, financial newsletters, credit-rating agencies, and other information producers can benefit from growing large. A caveat is that individual members can continue to monitor (and trust) each others as the organization grows large. If not, "free rider" problems will crop up, and it may not be beneficial to grow beyond a certain size because of the difficulty of implementing effective internal controls.

Asset Transformation as a Natural Monopoly: Now consider an asset transformer like a bank. It borrows money from depositors and makes loans. Its advantage in being large comes from two sources.[5] First, suppose multiple depositors are needed to finance a single bank borrower and the borrower's creditworthiness has to be established through costly credit analysis. Then having a bank perform this credit analysis once conserves screening resources compared to a situation in which all the depositors engage in costly screening of the borrower. That is, a bank eliminates duplicated screening. Second, the depositors' payoff is a debt contract, it is a concave function of the bank's payoff as shown on the next page.

Because the depositors' payoff is concave, they behave as if they are risk averse. Hence, they can be made better off by reducing the risk they face, and the benefit of this is a lower interest rate on deposits. The bank can do this by diversifying its risk across many different borrowers. And, because the benefit of diversification keeps growing with size, the bank is a natural monopoly.

5. The discussion below is based on a model developed by Diamond (1984).



How Banks Can Help to Make Nonbank Financial Contracting More Efficient

We have spent quite some time examining the flow of services that banks and other F.I.s produce. These services essentially take the form of intermediating in different ways between the users and providers of capital and of reducing their costs of exchanging capital. It has been suggested that banks not only permit the capital that flows through them to be exchanged at lower cost, but they also lower the cost of capital exchange between other parties.[6]

To understand this argument, let us examine the role of bank loans in a borrowing organization's information process. It is worthwhile to draw a distinction between *inside* and *outside debt*. Inside debt is defined as a contract in which the creditor has access to information about the borrower not otherwise publicly available. The creditor may even participate in the borrower's decision process. This could be achieved, for example, by the creditor having representation on the borrower's board of directors. Bank loans are inside debt. By contrast, outside debt is defined as publicly traded debt in which the creditor depends on information about the borrower that is publicly available. Commercial paper and publicly traded corporate bonds are examples of outside debt.

Bank loans offer a special advantage in this regard. They are usually of short maturities. This means they must be periodically renewed. These renewals are accompanied by bank evaluation of the borrower's ability to meet fixed payment obligations. Thus, if the bank renews a borrower's loan, it sends a positive signal about the firm to its other creditors. Note that credibility of this signal derives from the fact that the bank "puts its money where its mouth is" when it renews the loan. Given this credible and positive signal, other higher-priority creditors find it unnecessary to expend their own resources to duplicate the bank's evaluation. Thus, bank loans help to reduce duplication in borrower evaluation by multiple creditors.[7]

Banks also have a cost advantage in making loans to depositors.[8] The ongoing history of a borrower as a depositor communicates valuable information to the bank about the borrower's cash management activities. This permits the bank to assess the

6. See Fama (1980).

7. The argument that banks can lower the contracting costs of other parties can also be found in Fama (1990). For empirical work that follows upon the study discussed in this section, see Lummer and McConnell (1989).

8. This has been suggested, for example, by Black (1975) and Fama (1980).

risks of loans to depositors and to monitor these loans at lower cost than other (competing) lenders. This consideration is particularly important in short-term loans that are rolled over because of the relatively more frequent borrower assessments. This hypothesis has empirical validity in the observation that most short-term debt is in the form of bank loans.

The Empirical Evidence: Banks Are Special

It turns out that there is some interesting empirical support for the theories we have presented thus far. The central question of empirical interest to us is whether bank loans are unique, that is, do they provide any special service with their lending activity that is not available from other lenders? To answer this question we can examine the stock price responses to announcements of bank loans and other types of debt such as private placements of debt and public debt issues. The empirical evidence is that there is a positive and statistically significant stock price response to a borrower's acquisition of a bank loan. Further, the positive market reaction is not common to *all* private debt placements. There is, for example, a negative stock price response to debt placed privately with insurance companies. These findings seem to suggest that bank loans are unique.[9]

To examine these results let us first look at Table 3.2, which gives the distribution of announcements of different types of debt contracts for NYSE and AMEX firms.

Although there is no noticeable pattern in bank loans through time, there are two interesting observations. First, privately placed debt has been declining through time. Second, among all privately placed debt (bank loans plus other privately placed debt), bank loans dominate to the tune of 68.38 percent.

TABLE 3.2 Distributions by Year of Announcements of Bank Credit Agreements, Privately Placed Debt, and Publicly Placed Straight Debt for a Random Sample of 300 NYSE and AMEX-Traded Nonfinancial Firms for the Period 1974–1983

Year of Announcement	Bank Loan Agreements	Privately Placed Debt	Public Straight Debt
1974	9	4	5
1975	11	7	13
1976	7	7	8
1977	8	7	4
1978	1	8	6
1979	8	1	9
1980	11	1	10
1981	9	1	9
1982	10	1	16
1983	6	0	10
Total	80	37	90

Source: James, C., "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics* 19, 1987, 217–235.

9. See James (1987).

In Table 3.3 we provide descriptive statistics for different types of debt.

As this table shows, firms using private placements and bank loans are on average smaller than firms using public offerings of debt. The average firm size in both the bank loan sample and the private placement sample is about 25 percent of the average firm size in the public debt sample. This evidence is consistent with the theory discussed thus far. Problems of moral hazard and particularly of asymmetric information can be expected to be more severe for smaller, lesser-known firms. Hence, banks have a greater relative contribution to make in resolving these problems in such firms. Not surprisingly then, we find that bank loans are the dominant source of debt financing for small firms.

Let us now see how the stock prices of borrowing firms react to the announcements of various forms of debt. This evidence is presented in Table 3.4.

The abnormal stock return here is defined in the usual fashion as the deviation of the realized rate of return from the expected rate of return given by the market model. That is, the abnormal stock return for firm j over day t is defined as

$$R_{jt} - (\hat{\alpha}_j + \hat{\beta}_j R_{mt})$$

where R_{jt} is the rate of return of security j over day t, R_{mt} is the rate of return on the market portfolio over the same period, and $\hat{\alpha}_j$ and $\hat{\beta}_j$ are the ordinary least squares estimates of the market model parameters for firm j.

The average abnormal stock return for bank loan agreements in Table 3.4 is positive and statistically significant at the 0.01 level. In addition, two-thirds of the abnormal stock returns are positive. The negative average abnormal stock return associated with the announcement of a public offering of debt is not statistically significant.

If the positive response to bank loan agreements results from some benefit of inside debt not unique to banks, then one would expect to observe a similar response to debt that is privately placed with insurance companies. However, as Table 3.4 indicates, the response to the announcement of privately placed debt is −0.91 percent, which is statistically significant at the 0.10 level. Moreover, the difference between the average abnormal stock returns of bank loan agreements and privately placed debt is statistically significant at the 0.01 level.

TABLE 3.3 Descriptive Statistics for Commercial Bank Loans, Privately Placed Debt, and Publicly Placed Straight Debt for a Random Sample of 300 NYSE and AMEX-Traded Nonfinancial Firms for the Period 1974–1983

	Type of Borrowing					
	Commercial Bank Loans (Sample Size 80)		*Privately Placed Debt (Sample Size 37)*		*Public Straight Debt (Sample Size 90)*	
Descriptive Measure	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Debt amount (millions of dollars)	72.0	35.0	32.3	25.0	106.2	75.0
Firm size (millions of dollars)	675	212	630	147	2506	1310
Debt amount/market value of common stock	0.72	0.46	0.52	0.25	0.26	0.15
Maturity of debt	5.6	6.0	15.34	15.0	17.96	20.0

Source: James, C., "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics* 19, 1987, 217–235.

TABLE 3.4 Average Two-Day Percentage Abnormal Stock Returns on the Announcement of Commercial Bank Loans, Privately Placed Debt, and Publicly Placed Straight Debt Offerings for a Random Sample of 300 NYSE and AMEX-Traded Nonfinancial Firms for the Period 1974 to 1983

Type of Event	Abnormal Stock Returns	Proportion Negative (Sample Size)
Bank loan agreement	1.93%	0.34 (80)
Privately placed debt	−0.91%	0.56 (37)
Public straight debt	−0.11%	0.56 (90)

Source: James, C., "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics* 19, 1987, 217–235.

It is possible that the differences in abnormal stock returns across different types of debt agreements could be due to systematic differences in maturity and purpose of borrowing, that is, the data may not indicate anything special about bank loans *per se*. To check this possibility, we would like to know the share price responses to the announcements of bank loans, private placements, and public debt offerings, all with the same characteristics. The evidence on this score suggests that differences in abnormal performance across these different sources of borrowing are not solely due to differences in the characteristics of the loan or differences in the characteristics of borrowers (such as size, for example). That is, the results are robust. The overall conclusion to be reached from this empirical evidence is that banks are special.

Ownership Structure of Depository Financial Institutions

Depository institutions have two types of ownership forms: stocks and mutuals. Agency theory predicts that ownership form has a significant effect on the incentives and the operating efficiency of the firm. In this section, we will review the theoretical bases for this prediction and also look at some empirical evidence.

Commercial banks are exclusively stockholder-owned. Mutuals are common among insurance firms, MSBs (mutual saving banks), and S&Ls (savings and loan associations), although many mutual S&Ls have converted into stockholder-owned organizations in recent years. We will proceed as follows. First, we will examine how mutuality affects the resolution of agency and other problems. Then, we will seek an explanation for why S&Ls were dominantly mutuals and why the recent wave of conversions to stock ownership. Finally, we will review some relevant empirical evidence.

Mutual Versus Stocks

The residual claimants in a mutual are customers. These are the policyholders of mutual life insurance companies, the depositors of MSBs, and the depositors of mutual S&Ls. For purposes of this discussion, we will limit ourselves to mutual S&Ls.

There are two key differences between a stock and mutual S&Ls. First, the owners of a stock S&L are its stockholders, whereas the owners of a mutual S&L are its depositors (and possibly its borrowers). Second, a stock S&L can increase its capital by selling common stock, whereas a mutual S&L cannot.

Consider the first difference. In a stock S&L, shareholders have a well-defined ownership right, which implies: (i) a claim to residual profits, (ii) a right to vote for the board of directors and change control of the organization, and (iii) a right to

dissolve the organization. On the other hand, in a mutual S&L, the ownership rights of depositors are much weaker. As for (i), depositors in a mutual are much more like creditors than shareholders since they cannot force the mutual to pay them more than the promised interest and principal on their claims. Although in principle depositors have ownership claims to the mutual's current earnings, these claims are not transferable, and the earnings can be retained indefinitely by the institution as net worth.[10] As for (ii), while mutual S&L depositors have voting rights, these are quite limited and are often signed over to management at the time of opening of accounts.[11] Finally, as for (iii), even though a depositor can withdraw his deposits and thereby partially liquidate the mutual fund,[12] depositors have had little incentive to do so because of deposit insurance, especially when interest rate ceilings bounded the return to depositors.[13]

Thus, it is imperative to distinguish between *de jure* and *de facto* ownerships in a mutual. The *de jure* ownership (legal ownership) rests with the mutual's customers. It is, however, largely vacuous. The *de facto* ownership [control of (i), (ii), and (iii)] rests with the managers and the government (which provides deposit insurance).

Of course, the inability of owners to completely control the institution—and the resulting agency problem—is encountered in stockholder-owned institutions as well. Both stock and mutual S&Ls are administered by managers whose goals may differ from the goals of the owners. However, the two types of S&Ls differ with regard to the ability of the owners to monitor managers. Stockholders have greater control over the activities of managers because control can be consolidated through the purchase of stock.

Agency Problems in Stocks and Mutuals

The above discussion suggests that agency problems in mutuals should be greater than those in stockholder-owned institutions. There are two ways in which we can measure the incidence of agency problems. First, we can examine whether managers in mutuals spend more—and therefore operate less efficiently—than managers in stockholder-owned firms. The increased spending may be due to excessive consumption of perquisites by managers, less efficient cost control, or other expense-preferring behavior. Note that such behavior also represents a tension between the two *de facto* owners of mutuals, managers and the government. Managers may prefer to inflate expenses, whereas the government prefers that the mutual reduce expenses and increase retained earnings since this improves the institution's safety and diminishes the liability of the deposit insurance fund. Second, we can ask whether mutual S&Ls have operated at output levels as efficient as those of stock S&Ls. In other words, do mutuals exploit scale economies as efficiently as stock S&Ls?

The empirical evidence sheds light on these questions. Many studies have shown that managers in mutuals exhibit expense-preference behavior relative to

10. Indeed, during 1966–1982, cash distributions to depositors were legally prohibited under FHLBB interest rate ceilings. See Masulis (1987).

11. This is achieved with the signing of perpetual proxies. These proxies can be revoked. However, disclosure requirements on the part of the S&L management are limited, the maximum number of votes a depositor can control is limited, there are restrictions on outside nominations to the board, and the board can eliminate a depositor's voting rights by simply redeeming his savings account. See Masulis (1987).

12. See Fama and Jensen (1985).

13. See O'Hara (1981).

those in stocks.[14] Moreover, other studies have found that mutuals operate at inefficient output levels relative to stocks. For example, mutual S&Ls have been found to expand deposits and loans beyond profit-maximizing levels.[15] Of course, such behavior could be motivated by a managerial desire to consume additional perquisites, so that this inefficiency could be the outcome of expense preference as well.

Another output inefficiency may be found in diseconomies of scope in mutuals. The larger the number of products the firm produces, the more complicated its management structure, and the more costly it is for owners to monitor management. Thus, managers may be tempted to expand the product offerings of their firms beyond the level at which economies of scope are maximized. There is empirical evidence that suggests that diseconomies of scope are greater in mutuals than in stocks.[16]

Choice of Ownership Structure by S&Ls

Earlier studies viewed mutual S&Ls as either cooperatives, with depositors and borrowers working for a common goal, or benevolent associations organized to encourage saving and home ownership.[17] This view was based partly on the observation that the first S&Ls were mutuals that served smaller depositors, leaving the larger ones to commercial banks and other institutions.[18] These early community-based cooperatives, which gathered deposits from the community and offered mortgages to community members, had simple operations. The fair degree of homogeneity in mortgages made it relatively easy to assess the value of the S&L's assets based on historical data. This was just as well since the absence of a *secondary market* for residual claims meant that existing and prospective owners could not rely on the information generated by capital market trading (and pricing) to assess the value of the mutuals' assets. For assets whose value is difficult to determine stock ownership is superior because the information generated by trading facilitates valuation.[19] Whereas the simplicity of the operation of S&Ls made mutuality an acceptable ownership structure, the elimination of the classic conflict between creditors (who prefer less risk) and stockholders (who prefer more) made mutuality the *preferred* structure for many S&Ls.[20] Moreover, the simplicity of the operation of S&Ls meant that managerial expertise was not a critical element in the success of S&Ls. In the early years, therefore, the S&L industry was dominated by mutuals run by managers who were not the most talented or efficient.

Over time, however, operation became more complex, and mutuals began to choose managers on the basis of expertise.[21] Moreover, the advent of deposit insurance eliminated the agency-cost-of-debt advantage of mutuals over stocks. Since their deposits are insured, depositors are indifferent to an S&L's risk-taking behavior. The agency cost of debt was essentially absorbed by the Federal Savings and Loan Insurance Corporation (FSLIC).

14. Deshmikh, Greenbaum, and Thakor (1982) make this theoretical prediction. Supporting empirical evidence can be found in Edwards (1977), Hannan and Mavinga (1980), and Smirlock and Marshall (1983).

15. See Akella and Greenbaum (1988).

16. See Mester (1991) for careful empirical documentation that stock S&Ls operate with an efficient output mix, whereas mutual S&Ls operate with significant diseconomies of scope.

17. See Hester (1968) and Brigham and Pettit (1969).

18. There are also theoretical models that suggest such a role for mutuals. See Rasmusen (1988).

19. See Fama and Jensen (1983).

20. See Meyers and Smith (1986).

21. See Masulis (1987).

Along with these developments came deregulation and an increase in competition. Mutual S&L managers have found it increasingly difficult to compete with their more efficient stockholder-owned counterparts. And their inability to augment institutional net worth through additional equity issues has made the competitive disadvantage worse. Thus, the benefits of mutuality to owners have diminished significantly. Furthermore, these increased competitive pressures mean that the probability of bankruptcy—and hence the probability of unemployment for the manager—due to inefficient behavior has increased. This means that any given level of perquisite consumption on the part of mutual managers is now more costly. Given that managers were optimally selecting their perquisites prior to deregulation, the implication is that perquisites consumption in mutuals must be *lower* after deregulation, as managers weigh the benefit of perks against the elevated probability of unemployment. Thus, the benefits of mutuality to managers have diminished as well. Combined with this is the positive incentive managers have to convert to stock ownership, since they usually benefit in the initial stock sale. The reason is that managers typically receive rights to purchase the new stock, which is usually underpriced (as in other initial public offerings). When the benefits of conversion outweigh the benefits of the new optimal (and lower) level of perquisites consumption, the S&L will convert from mutual to stock.[22] This could explain the increased number of conversions that have been witnessed in recent years,[23] as the stockholder-ownership structure has become the preferred mode for both owners and managers.

The Borrower's Choice of Finance Source

We have seen that a borrower has access to a wide array of credit sources. How does he decide which source to approach? In Figure 3.2 below, we have sketched a hierarchy of financing sources that explains the borrower's choice based on his own attributes and the resulting demand for intermediation services.[24] The borrower's financing choice in this figure tracks a typical firm's "lifecycle."

When a firm is very young, it has two striking characteristics. First, the entrepreneur in charge may be unsure of his own management expertise, so that approaching a financial intermediary that can provide this expertise is beneficial. Second, the

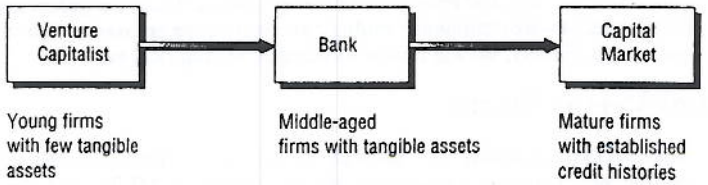


FIGURE 3.2 Hierarchy of Financing Sources

22. Mester (1991) arrives at this explanation based on her empirical analysis.

23. To convert from mutual to stock, the S&L must sell stock publicly through a standby rights offering to depositors and management, who are the eligible subscribers. The conversion plan must be first approved by two-thirds of the S&L's board of directors. If approved, it must be ratified by two-thirds of the depositors. Upon ratification, the stock can be offered to eligible subscribers, and if it is not fully subscribed, the unsubscribed portion must be sold to the public.

24. This discussion is based in part on Diamond (1989), and Chan, Siegel, and Thakor (1990). See also Boot and Thakor (1997).

borrower has few tangible assets to offer as collateral. As we will see in Chapter 5, collateral is useful in controlling moral hazard whereby borrowers either stint on effort or take excessive risks. In the absence of collateral, the lender could use equity participation as a way of addressing moral hazard. Thus, it is in the borrower's interest to seek a lender who can take an equity position and thus be able to offer capital at a "reasonable" price. Both factors suggest that such firms should go to venture capitalists.

As a firm grows and acquires tangible assets, it becomes capable of offering collateral to mitigate moral hazard. Banks, which are prohibited in the United States from taking equity positions, can now lend to such borrowers because they can offer collateral to secure their debt. Of course, all moral hazard will not be eliminated by collateral, so that there will be an important role for bank monitoring. Moreover, bank loans tend to be of short maturities, thereby generating periodic information through reassessments of the borrower. This information is reflected both in the bank's decision to renew/terminate the loan as well as in the new contract terms offered, in combination with information produced by rating agencies. This helps to reduce duplication in information production by *other* creditors of the firm, thereby diminishing overall contracting costs. The firms in this group find it better to go to banks than to venture capitalists because banks can fund their loans with insured deposits, whereas venture capitalists cannot; hence, the borrower is able to obtain a loan at a lower price.

Finally, when the firm is well-established and mature, it has a good track record for repaying its debts. This reputation can be valuable because it permits the firm to borrow at preferential rates. By taking undue asset risks, the borrower stands to lose this reputation, and thus has an incentive to limit risk-taking. Consequently, bank monitoring to combat moral hazard is less important for such borrowers, and this permits them to directly access the capital market where borrowing costs are lower; capital market access would mean that the borrower would not have to pay the bank its intermediation rents. Of course, such firms still confront problems of asymmetric information[25], so that *nondepository* financial intermediaries such as investment banks (or credit-rating agencies) play an important role in the transfer of capital from investors to such firms. This is because they make information about firms available to investors at a lower cost than they could acquire themselves. It is interesting to note that as one moves from left to right in the financing hierarchy of Figure 3.2, the intermediation services provided decline and so does the cost of credit. The venture capitalist provides financing, monitoring, and management expertise; the bank provides financing and monitoring; and the capital market provides mainly financing. Of course, this discussion is not meant to suggest that these financing sources are mutually exclusive. For example, borrowers often access the capital market for commercial paper and use banks to provide loan commitments to back up these commercial paper issues.

Blurring Distinctions Between Bank Loans and Capital Market Financing: Transaction and Relationship Loans

Although in our earlier discussion, we have characterized capital market and bank financing as distinct but sometimes overlapping choices, in recent years the distinction between these two sources of financing has become increasingly blurred. For

25. See, for example, Myers and Majluf (1984).

example, banks made syndicated loans in which multiple banks participate, and these loans are often traded in a manner similar to capital market trading. Banks make mortgage and credit card loans and then package them into portfolios, issue securities against these portfolios and sell these securities in the capital market where they are traded. This is called securitization and will be discussed in more detail in Chapter 9.

Of course, banks also make loans where they add considerable unique value and the loans are not traded. Examples are small business loans where the bank-borrower relationship has value.

Research in banking has examined the difference between loans by classifying bank loans as transaction loans and relationship loans.[26] Transaction loans include loans like credit card and mortgage loans. There is little monitoring by the bank and the loans can be repackaged and traded. The bank's value added is limited mostly to its credit analysis and standardized credit analysis before credit is extended.

Relationship loans are those where the bank generates additional value by learning about the borrower through its relationship with the borrower and providing business advice. Relationship loans offer numerous other advantages related to attenuating moral hazard and private information problems. These will be discussed in Chapter 6.

Another aspect of relationship lending that has only recently begun to be explored is that it creates the potential for *differences of opinion*. For example, a bank may judge a relationship loan to be creditworthy, but its judgment may be based on a lot of "soft," nonverifiable information. Such loans may find it difficult to obtain direct capital market financing if investors have a different (collective) opinion about the creditworthiness of the loan. In such cases, a bank—backed by sufficient capital—can act as a "beliefs bridge" between depositors/investors and borrowers and raise deposit financing to fund the relationship loan. The bank's reputation/credibility is reliably processing soft information and this may convince depositors to extend funding they otherwise may not have. This would be another contribution of banks to relationship loans.[27]

Thus, bank loans span a continuum from relationship loans at one end to transaction loans at the other. Relationship loans are the most different from capital market financing. Transaction loans are the most similar to capital market financing.

Conclusion

The process of financial intermediation is of central importance to the functioning of a modern economy. Some of the important conclusions to be drawn from our discussions are covered briefly below.

First, regulation of banks and the *raison d'etre* for the existence of banks are intertwined. Regulation is not solely the outcome of a political agenda that is separate from the reasons why banks exist. To make banking a viable business in which there is public confidence, some form of regulation is necessary. We also discussed how this regulation then becomes a component of monetary policy. Second, the incentive problems that banks and nondepository financial intermediaries resolve are such that there are natural benefits to size. Diversification can reduce incentive costs in

26. This characterization was provided by Boot and Thakor (2000). See also Rajan (1993) and Sharpe (1992) for models of relationship lending. Boot (2000) provides a review.

27. See Coval and Thakor (2005) and Song and Thakor (2006).

contracting among unequally informed agents, and information reusability is greater in larger intermediaries. Hence, financial intermediaries can derive economic benefits from being large. Third, inside (privately placed) debt has some inherent advantages over outside (publicly traded) debt because of superior access to information about the borrower that the former provides. Bank loans are inside debt. However, even within the class of contracts qualifying as inside debt, bank loans are special. The reaction of a borrowing firm's stock price to the announcement of a bank loan agreement is more favorable on average than the stock price reaction to the announcements of other forms of inside debt. Fourth, the choice of organizational form—mutual versus stock—by a depository institution depends on the interaction between a variety of factors that include differences in the efficiency with which agency problems are resolved within mutuals as opposed to stocks, the competitive environment, and the relative advantage a stockholder-owned firm has in raising capital and having complex assets priced in the capital market. This explains the initial prevalence of mutuality among thrifts and the recent trend of conversions of mutuals into stock. Finally, there is a natural hierarchy of financing sources. In its earliest phases of development, a firm has the greatest advantage in seeking venture capital, due to the (unique) ability of the venture capitalist to assist in management. At the next stage, when early survival has been accomplished, bank loans are preferred. Although banks do not assist in management to the extent that venture capitalists do, the monitoring provided by banks is of value to firms at this stage when they are still relatively small or medium-sized. Bank monitoring helps to control incentive problems within the borrowing firm. Moreover, bank loans tend to be of short maturities, thereby generating periodic information through reassessments of the borrower. This information, as well as that produced by nondepository financial intermediaries such as credit-rating agencies, helps to reduce duplication in information production by *other* creditors of the firm, and thus reduces overall contracting costs. Finally, large firms go directly to the capital market for outside debt. Bank monitoring is of lesser marginal value to such firms. However, such firms still confront problems of asymmetric information,[28] so that *nondepository* financial intermediaries such as investment banks (or credit-rating agencies) play an important role in the transfer of capital from investors to such firms. This is because they make information about firms available to investors at lower cost than they could acquire themselves.[29]

What are the implications of our analysis for *market efficiency*? Clearly, if the capital market were strong-form efficient even without financial intermediaries, the role for financial intermediaries would be extremely limited; they would at best provide some minor transactional services like "lot-breaking" of securities, that is, buying large denomination securities and selling smaller denomination claims against such securities to investors with wealth constraints. However, the theoretical and empirical results discussed in this chapter suggest two conclusions. First, given the pervasive problems of private information and moral hazard, it is reasonable to expect that credit markets are no more than semistrong form efficient, so that financial intermediaries have an important role to play in resolving information-based problems. Second, the informational efficiency of credit markets is *enhanced* by financial intermediaries, since they possess privileged financial information that is then learned by others who observe bank-borrower transactions.

28. See Ramakrishnan and Thakor (1984) and Giammarino and Lewis (1988).
29. See, for example, Diamond (1989).

Review Questions

1. Explain how a bank evolves from a primitive goldsmith and the roles played by asymmetric information and moral hazard in this evolution.
2. Can banking ever become completely deregulated? Why or why not?
3. What do we mean by a "hierarchy of financing sources"? What determines a borrower's choice of financing source?
4. Can you shed light on the following facts and explain their possible interrelationships?
   a. Commercial paper issues by nonfinancial corporations in the U.S. have grown sixfold in the last 20 years.
   b. Large money center banks are turning increasingly to "middle market" borrowers (that is, those with loan requests between $5 million and $200 million).
   c. Securitization has grown rapidly.
5. What is the difference between a "stock" and a "mutual"? Explain the differences in the resolutions of agency problems for these two types of organizations.
6. It has been said that the health of a nation's banking system is inversely related to the speed and efficiency of information flows in the economy. Explain.
7. In what way are banks "unique"? What is the empirical evidence on this issue?
8. What are the economic incentives for financial intermediaries to grow large?
9. How do banks help to make nonbank contracting more efficient?
10. Given below is an excerpt from "A Friendly Conversation." Comment critically on it.

    Moderator: Fine, but as long as you have fractional reserve banking, you're never going to eliminate the possibility of withdrawal risk altogether.
    Appleton: That's why you have a lender of last resort, Mike.

11. How does monetary policy affect the (short-term) growth path of an economy?
12. What are the differences between transaction and relationship loans and what is the relevance of the distinction?

Appendix 3.1 The Formal Analysis of Large Intermediaries

The Model Based on Ramakrishnan and Thakor (1984): Suppose we have assets whose owners wish to attract capital. However, there is asymmetric information about the values of these assets; the owner of each asset knows more about the value than others do. As we saw in Chapter 1, this can lead to market failure if the appropriate signals are unavailable to firms. Now suppose there are some individuals who specialize in producing information about firms at a cost. Let us imagine that there are groups of individuals, with each group specializing in producing information about a particular industry or a particular firm. The cost to an individual of producing this information is $c > 0$ and each individual is risk averse, with a utility function of $U(\bullet)$ defined over monetary wealth, that is $U(\bullet)$ is increasing and strictly concave. We assume that c is a nonmonetary cost to the information producer (i.p.); it does not

figure in his utility over wealth. Moreover, it is incurred only if the i.p. actually produces information about the firm he specializes in. Also, each i.p. has a minimum level of expected utility, a $\bar{U}$ that must be guaranteed by his compensation package for producing information, or he will work in an alternative occupation.

Now suppose that the firm that wishes to attract capital (or the investor who wants to decide whether he should invest in a particular asset) approaches an i.p. directly to produce information about it and release it to the market, that is, the i.p. plays the role of a rating agency. If the i.p. is just paid a fixed fee, we have a moral hazard problem in that he will avoid actually producing information, thereby saving himself the effort-related cost c. He will simply make a quick guess, collect his fee, and send the firm on its way. Investors will recognize this and the firm's price will not move. The firm will have wasted its money.

Compensation Contracts of Individual Information Producers: But suppose the firm is able to monitor the i.p. to discover something about whether he actually invested c. This monitoring produces a signal that tells the firm about the i.p.'s effort. However, this signal is noisy. Even if the i.p. invests c in information production, the signal says that he did only with probability p. With probability $1 - p$, the signal is erroneous and indicates that the i.p. did not produce information. If the i.p. did not produce information, then the signal says that he did with probability q and that he did not with probability $1 - q$. We assume $p > q$, so that the signal is informative. Now let the i.p.'s compensation be as follows: pay him \$H if the signal says he produced information and \$L if it says he did not, with $H > L$.[1] If the i.p. does produce information, he gets an expected utility of

$$EU(\text{produce information}) = pU(H) + (1 - p)U(L) - c. \qquad [3.13]$$

If he does not produce information, he gets an expected utility of

$$EU(\text{does not produce information}) = qU(H) + (1 - q)U(L). \qquad [3.14]$$

If investors are to believe that the i.p. is credible, his compensation schedule should be incentive compatible (should induce the i.p. to invest c). That is,

$$pU(H) + (1 - p)U(L) - c \geq qU(H) + (1 - q)U(L). \qquad [3.15]$$

It also will be necessary to make sure that the i.p. is willing to work for the firm. This requires that

$$pU(H) + (1 - p)U(L) - c \geq \bar{U}. \qquad [3.16]$$

We can solve (3.15) and (3.16) to come up with H and L. We can show that in equilibrium (3.15) and (3.16) should hold as equalities, that is, treating them as equalities leads to a solution that minimizes the expected cost for each firm. To illustrate, suppose $U(x) = \sqrt{x}$ for any number x, $\bar{U} = 20$ (for simplicity), $p = 0.8$, $q = 0.2$ and $c = 10$. Solving (3.15) and (3.16) as a pair of simultaneous equations with these numbers, we get $H = 10{,}000/9$ and $L = 10{,}000/36$. The i.p. earns an expected utility of exactly 20. The expected cost of information production for each firm is $0.8\,H + 0.2\,L = 944.44$ approximately.

1. If such a compensation scheme is successful in inducing the i.p. to produce information, then it is not time consistent because everybody knows he has produced information and it is pointless to pay him less when an error-prone signal says he did not. We'll ignore this problem here.

The Solution With an Intermediary: Now suppose that there are two i.p.s, each like the i.p. in the preceding analysis, who coalesce and form a financial intermediary of two i.p.s. Each still deals with a separate firm. However, they now pool their payoffs to avail of diversification benefits. We assume that because the i.p.s are cooperating, they can costlessly observe each other's actions. This means neither i.p. has to be concerned about his partner free-riding off his effort. So now each i.p.'s compensation becomes

$2H/2 = H$ if both signals are favorable
$(H + L)/2$ if only one signal is favorable
$2L/2 = L$ if both signals are unfavorable

Assuming that signals across firms are uncorrelated, the probabilities of different compensations for each i.p. are given in the following table.

TABLE 3.5 Probabilities of Compensations

Probability of Compensation	Compensation of Each i.p.
p^2 if both i.p.s produce information and q^2 if both do not	H
$2p(1 - p)$ if both i.p.s produce information and $2q(1 - q)$ if both do not	$(H + L)/2$
$(1 - p)^2$ if both i.p.s produce information and $(1 - q)^2$ if both do not	L

Note that both i.p.s will act in concert. The firms that give them compensation contracts realize that the rules of the game have changed. They must now solve the following pair of simultaneous equations.

$$p^2U(H) + 2p(1 - p)U\left(\frac{H + L}{2}\right) + (1 - P)^2U(L) - c$$
$$= q^2U(H) + 2q(1 - q)U\left(\frac{H + L}{2}\right) + (1 - q)^2U(L) \qquad [3.17]$$

and

$$p^2U(H) + 2p(1 - p)U\left(\frac{H + L}{2}\right) + (1 - p)^2U(L) - c = \bar{U} \qquad [3.18]$$

Generally, the solution to this will be different from the previous solution. Suppose, however, that firms continue to use the old contracts where $H = 10.000/9$ and $L = 10.000/36$. It can be checked in this case that (3.17) is satisfied exactly and that the left-hand side of (3.18) is about 20.43. That is, each i.p. in the financial intermediary enjoys a higher expected utility than he did before. Note that the expected cost of having information produced for each firm will be exactly the same as before. Thus, the formation of a financial intermediary makes i.p.s better off if firms do not alter their contracts. Of course, firms may wish to write different contracts to remove the excess utility enjoyed by the i.p.s. In this case, expected information production costs of firms are lowered.

The reason why the formation of an intermediary helps is diversification. By pooling their payoffs, the i.p.s are able to reduce individual risks. This means that they can increase their expected utility and if at least some of the benefit of this increased utility is shared with the firms they are screening, the cost of information production will also decline.

The Desirability of a Very Large Intermediary: This argument can be taken to the limit. Suppose the financial intermediary becomes infinitely large. Then, by the law of large numbers (roughly speaking) the probabilities become actual fractions. That is, if all i.p.s produce information, the intermediary knows that exactly 80 percent of them will get H each and 20 percent will get Leach. Thus, the intermediary knows that its payoff will be

$$0.8\mathrm{H} + 0.2\mathrm{L}$$
$$= 0.8\left(\frac{40000}{36}\right) + 0.2\left(\frac{10000}{36}\right) = 944.44$$

per i.p. with probability one. Since the financial intermediary itself can monitor its own members, it does not have to worry about moral hazard. Thus, it can promise each of its member i.p.s a *fixed* payment of 944.44, knowing that even though on any given i.p., it could receive either more or less than this amount, the random fluctuations around 944.44 will cancel out for the intermediary as a whole. Thus, each individual i.p.'s expected utility in this intermediary is $U(944.44) - 10 = 20.73$, which is higher than with the two-i.p. intermediary passes along this gain to the firms it screens, then *information production costs are lowest with a very large intermediary.*

That is, we have shown that a *diversified information broker* can lower the cost of information production and hence the cost of exchanging capital. Once again, the pivotal function served by a financial intermediary is that of providing a more efficient resolution of informational problems.

Diversification in this model is achieved by letting each i.p. within the intermediary share the risk in the compensation of every other member i.p. That is, as we add to the size of the group, each individual compensation risk is shared by an increasing number of i.p.s. Due to the risk aversion of the member i.p.s, such diversification helps to improve welfare.[2] We shall call this "diversification by sharing risks." Another type of diversification is "diversification by adding risks."[3] In this case, a single i.p. bears 100 percent of N independent risks, with diversification occurring as N increases. This is quite different from the first form of diversification because the total wealth of the i.p. is growing as he adds more risks. That is, instead of spreading a given amount of wealth over a larger number of independent gambles, we are spreading an increasing amount of wealth over a larger number of independent gambles. Noble laureate Paul Samuelson (1963) has called such diversification "the fallacy of large numbers," because it is not generally true that, for all risk-averse utility functions, the individual's risk aversion toward the N^{th} independent gamble is a decreasing function of N. In other words, while a risk-averse individual would wish to take advantage of the low number of large numbers to spread a *fixed* amount of wealth over an increasingly large number of independent gambles, he would not necessarily wish to achieve such diversification at the expense of exposing an increasing amount of his wealth to the gambles. However, there are sufficient conditions involving restrictions on utility functions that such diversification is beneficial.

2. An important assumption in our analysis is that the i.p.s within the intermediary can monitor each other costlessly. Millon and Thakor (1985) show that if such monitoring is impossible, then by letting i.p.s coalesce and engage in payoff-pooling, we raise information production costs. They also show, however, that if the values of firms depend on a common, systematic element, as well as on idiosyncratic factors, then information sharing within the intermediary can lead to on overall lowering of information production costs.

3. This is considered by Diamond (1984).

References

Akella, S. Rao, and Stuart I. Greenbaum, "Savings and Loan Ownership Structure and Expense-Preference," *Journal of Banking and Finance* 12, 1988, 419–437.

Allen, Franklin, "The Market for Information and the Origin of Financial Intermediation," *Journal of Financial Intermediation* 1–1, May 1990, 3–30.

Black, Fisher, "Bank Fund Management in an Efficient Market," *Journal of Financial Economics* 2, 1975, 323–339.

Boot, Arnoud, "Relationship Lending: What Do We Know?" *Journal of Financial Intermediation*, 2000.

Boot, Arnoud, and Anjan V. Thakor, "Can Relationship Banking Survive Competition?" *Journal of Finance* 55–2 April 2000, 679–714.

Boot, Arnoud, and Anjan V. Thakor, "Financial System Architecture," *Review of Financial Studies*, Fall 1997, 693–733.

Boyd, John, and Edward Prescott, "Financial Intermediary Coalitions," *Journal of Economic Theory* 38, April 1986, 211–232.

Brigham, Eugene F., and Richard R. Pettit, "Effects of Structure on Performance in the Savings and Loan Industry," in *Study of the Savings and Loan Industry* (Irwin Friend, Ed.), Federal Home Loan Bank Board, Washington, DC, 1969.

Campbell, Tim, and William Kracaw, "Information Production, Market Signalling, and the Theory of Financial Intermediation," *Journal of Finance*, September 1980, 35–4, 863–882.

Chan, Yuk-Shee, Daniel Siegel, and Anjan Thakor, "Learning, Corporate Control and Performance Requirement in Venture Capital Contracts," *International Economic Review* 31–2, May 1990, 365–381.

Coval, Josh, and Anjan V. Thakor, "Financial Intermediation as a Beliefs-Bridge Between Optimists and Pessimists," *Journal of Financial Economics* 75–3, March 2005, 535–570.

Deshmukh, Sudakar, Stuart Greenbaum, and Anjan Thakor, "Capital Accumulation and Deposit Pricing in Mutual Financial Institutions," *Journal of Financial and Quantitative Analysis* 17, December 1982, 705–725.

———, "Reputation Acquisition in Debt Markets," *Journal of Political Economy*, 97–4, 1989, 828–862.

Diamond, Douglas, "Financial Intermediation and Delegated Monitoring," *Review of Economics Studies* LI, July 1984, 393–414.

Edwards, Franklin R., "Managerial Objectives in Regulated Industries: Expense Preference Behavior in Banking," *Journal of Political Economy* 85, 1977, 147–162.

———, "Contract Costs and Financing Decisions," *Journal of Business*, 63–1–2, January 1990, S71–S91.

CHAPTER • 4

Major Risks Faced by Banks

"Bets on the directions of interest rates are like the little girl from the nursery rhyme with the curl on her forehead. When they are good, they can be very, very good, but when they are bad, as NCNB Corp. is now finding out, they can be horrid."

Kelley Holland: *American Banker*, March 20, 1990

Glossary of Terms

OTS: Office of Thrift Supervision. This is a national regulatory agency for the thrift industry.

Zero-coupon bonds: Bonds that pay no coupon, so that the entire repayment to bondholders is at maturity.

Immunization: The act of insulating the institution from interest rate risk.

Going Long: Purchasing a security.

Going Short: Selling a security without owning it.

Introduction

Risk is endemic to business but central to banking. What precisely do we mean by risk? In the context of business, risk is the distillate of randomness in the process by which earnings are generated. This randomness may be avoidable, in large part, in which case the risk is voluntarily accepted, perhaps even sought, as routine business decision; hence a "businessman's risk." Alternatively, the risk may be unavoidable, as in the case of a *force majeure* or an "act of god," in which case the only protection is to seek outside insurance or to exit the industry. The risks in business are as diverse as

life itself. The businessman faces possible losses owing to flood, plague, fire, machine failure, worker alienation, sabotage, war, or capricious acts of government that destroy or appropriate property (sovereign risk). Shoe stores as well as financial intermediaries face all of these risks, but risks of the avoidable variety define the business of banking.

It is important to bear in mind that risk is *not* due to *variability per se*, but rather due to *uncertainty*. In an *ex post* sense, we often use the terms variability and uncertainty synonymously. However, in an *ex ante* sense, the two are quite distinct. We can have a cash flow, for example, that is known *for sure ex ante* to be 1, −100, 1,000, and 0 in years 1, 2, 3, and 4, respectively. This cash flow has a very high intertemporal variability, but has *no* risk. By contrast, a cash flow that can be either +1 or −1 with equal probability in each of the next 4 years has less intertemporal variability but more risk.[1] Risk, then, is related to uncertainty or lack of predictability.

The Source of Business Risk

What kinds of risks do banks face? To address this question, it is important to note that banks are essentially no different from other firms when it comes to the *raison d'etre* for being exposed to risk. A bank's shareholders, or the shareholders of any other firm for that matter, bear risk when the economic nature of the firm's "assets" is somehow different from that of its "liabilities."

Consider a steel fabrication company in Figure 4.1 below.

In Figure 4.1, the risk to the fabricator's shareholders arises primarily from the fact that the prices of raw steel and fabricated steel do not move in perfect unison. This exposes the fabricator's profit margin to random fluctuations and creates risk for its shareholders. Note that this risk comes from a mismatch on the fabricator's "balance sheet." Its liability (what it owes its suppliers for raw steel) is of a different nature from its "assets" (the fabricated steel it sells to its customers) because the prices of raw and fabricated steel are not perfectly correlated.

Now suppose the fabricator purchases its raw steel in Japan, paying its suppliers in Japanese yen, and sells fabricated steel in the United States, receiving dollars from its customers. In this case, we see that the fabricator's balance sheet is even more mismatched because of the different currencies involved. Consequently, its

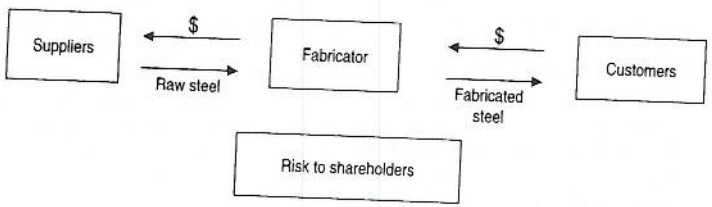


FIGURE 4.1 Risks Faced by a Domestic Steel Fabricator

1. For the havoc caused by not distinguishing between variability and risk, see Sprenkle and Miller (1980).

shareholders are exposed to even more risk. In particular, they face currency risk (due to the lack of perfect correlation between movements in the yen and the dollar) in addition to the price risk they faced earlier.

In general then, *mismatches* imply risks. This is a notion familiar to us from Chapter 2. Qualitative asset transformation involves mismatching the two sides of the balance sheet and, hence, creates risk. What are the major mismatches for banks? These are described in Figure 4.2.

In Figure 4.2 we see that a typical bank's assets (e.g., loans) and liabilities (e.g., demand deposits) are mismatched along three dimensions. First, the assets usually involve greater credit risk than the liabilities, i.e., the bank's claim against the borrower is riskier than the depositor's claim against the bank. Second, the assets are usually of longer maturity than the liabilities. For example, a loan may have a 1-year maturity, whereas demand deposits are withdrawable on demand (zero maturity). This creates interest rate risk. Third, a bank's liabilities are usually more liquid than its assets, i.e., a depositor is able to withdraw his deposits without notice, whereas the bank cannot call back a performing loan at-will and the loan may also not trade in an active market. This creates liquidity risk. We shall now discuss each of these risks in more detail.

Credit, Interest Rate, and Liquidity Risks

1. Default or Credit Risk: This is the risk that a party with whom you contract fails to fully discharge the terms of the contract. For a bank, this is the risk that a borrower fails to make the contractual payment on a timely basis. This kind of risk is central to virtually all rental transactions, and as in the case of almost all insurance contracts, moral hazard is a key element in default risk.

The avoidability of default risk has two aspects. Banks can choose assets with little or no default risk, such as government securities or the debt of triple-A rated borrowers. Such a strategy, however, may provide a return only slightly, if at all, greater than the bank's cost of borrowing, and such a low (albeit relatively safe) profit margin may be unattractive to the bank.

Given that the bank chooses assets with substantial default risk, its ability to control default risk derives from its ability to resolve moral hazard and other

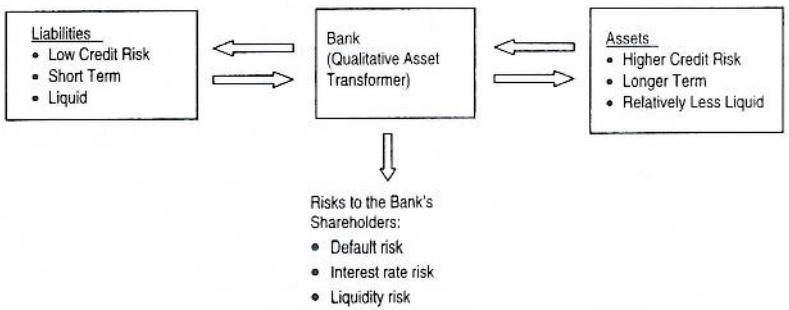


FIGURE 4.2 Major Mismatches for Banks

informational problems. In our earlier discussion in Chapters 2 and 3, we argued that banks enjoy a special advantage in screening and monitoring borrowers. However, it is virtually impossible to monitor a borrower so closely that default (or credit) risk can be completely eliminated.

There are two sources of default risk: cash flow variations beyond the borrower's control (physical hazard) and moral hazard. As for the first source of default risk, the bank's role as a financial intermediary is to *screen* the borrower so that it can accurately assess the risk it is taking in lending. This involves an analysis of the borrower's financial statements and other relevant financial and operating information about the borrower. In this capacity, the bank only assesses the risk but does not bear it, that is, it is acting as a pure broker. Thus, its role is similar to that of a bond rating agency or an investment banker. Our discussion in Chapter 3 suggests that large (diversified) information brokers can motivate their members to produce information at lower cost than is possible without intermediation. Thus, the bank should be able to efficiently generate information about default risk stemming from cash flow variations beyond the borrower's control.

As for moral hazard, the bank's monitoring capability is important. As we will see in some detail in the next chapter, the borrower has an incentive to take actions after taking a (risky) loan that increase the bank's risk exposure. This is why covenants are included in loan contracts to restrict the activities of borrowers. However, bank monitoring of borrower compliance with these covenants is important to control moral hazard. Thus, the efficiency with which the bank performs its basic functions as an FI is a key determinant of its own credit risk exposure. Moreover, loans are subject to management as a portfolio. A bank can control its default risk by holding in its asset portfolio many loans with imperfectly correlated prospects and thereby diversifying across loans.

2. **Interest Rate Risk:** This risk derives from variation of market prices. If the firm's assets and liabilities are traded, they are subject to being revalued by the market. Any such revaluation, due to changes in either the level or structure of interest rates, is described as interest rate risk. Let's consider a simple example. Suppose a bank makes a 2-year, $1 million loan for which it charges 10 percent interest. It faces the choice of financing the loan with a 2-year deposit at 9 percent per annum, or with a 1-year deposit at 8 percent per annum. The former choice will result in $10,000 in *certain* interest earnings for each of the 2 years. However, if the bank chooses the 1-year financing, it will earn $20,000 in year 1, but its earnings in year 2 will depend on the currently unknown 1-year interest rate that will prevail a year from now. Should the 1-year rate remain unchanged, the bank will enjoy a second year of earning $20,000. And if the 1-year rate were to fall to 5 percent, management will do even better and record second-year earnings of $50,000. But interest rates rise too, as the S&L industry discovered to its chagrin in 1980–81, and should the 1-year rate rise to, say, 12 percent, the bank will sustain a loss of $20,000 in year 2. This example illustrates both the substance of interest rate risk and its discretionary aspect. The risk could have been avoided with the choice of 2-year financing, assuming, of course, that 2-year financing was available. If not available, or if available only at a rate exceeding 10 percent, the bank need not have offered the borrower a 2-year fixed-rate loan.

Another aspect of interest rate risk, from the standpoint of the bank, is *prepayment risk*. This risk arises from the borrower's *option* to prepay. If interest rates rise, there will be no prepayment. But if interest rates fall sufficiently after the loan has been taken, the borrower is likely to prepay the loan by taking advantage of refinancing at a lower rate.

3. **Liquidity (Withdrawal) Risk:** This is the risk that an asset owner (seller of a house or a borrower selling its indebtedness) will not be able to realize the full value of that asset at the time a sale is desired. In banking, the liquidity risk faced by a borrower is that the lender may choose *not* to renew a loan that a borrower wants to renew. Similarly, the liquidity risk faced by a bank is that depositors may unexpectedly withdraw their deposits and the bank may be unable to replace them without impairing its net worth. This risk applies symmetrically to borrowers in their relationship to banks, and to banks in their relationship to depositors. The most extreme manifestation of liquidity risk is that the seller of the asset is simply *unable* to sell the asset at any price. In credit markets, this phenomenon is known as *credit rationing*, whereby a borrower is refused credit regardless of the price it is willing to pay. We shall have more to say about credit rationing in Chapter 6, but suffice to say this phenomenon has long perplexed economists in particular because it indicates an apparent suspension of price as the arbiter of allocations.

Liquidity risk has yet another interpretation. Asset markets vary widely in their development and level of activity. At one extreme, we have flea markets for "one-of-a-kind" antiques of dubious authenticity. At the other we have 24-hour around-the-world markets for currencies and government debt in which large quantities are traded at relatively low cost. More primitive and less active markets are typically characterized by large bid-ask spreads, where the *bid-ask spread* is defined as the difference between the price at which one can buy a security and the price at which one can sell it at the same place and time. For example, you can *buy* a Treasury bill at an *ask* of $98½ and *sell* it at a *bid* of $98¼, in which case the bid-ask spread is $98½ − $98¼ = $¼. Bid-ask spreads range from small fractions of a percent of the asset value for actively traded assets, to 6 or 7 percent for residential property. Still larger bid-ask spreads hold for infrequently traded, heterogeneous, and hard-to-value objects. Bid-ask spreads are the cost of simultaneous purchase and sale of an asset, and reflect the liquidity in asset markets.

Illiquid assets are those for which "full value" is not readily realizable. That is, time and effort are required to realize the full value of an asset that is relatively illiquid.[2] Hence, a bank holding illiquid assets can find itself unable to redeem its liabilities on short notice, and the problem of managing the balance sheet against this eventuality is referred to as liquidity or cash management (cash is the asset with liquidity *par excellence*). The central bank, with its capacious lender-of-last-resort facility, was created to address those instances when the bank, having sound albeit illiquid assets, is unable to meet its withdrawals. The central bank provides the bank with crisis-avoiding liquidity by lending to the bank against its illiquid but otherwise presumably sound earning assets. Indeed, the central bank was designed to socialize a portion of the bank's liquidity problem.

In the remainder of this chapter we shall address interest rate and liquidity risks in greater detail. Default risk will be considered in Chapters 5 and 6, where lending will be the focus. The rest of this chapter is organized as follows. First,

2. Were all assets perfectly liquid, there would be no role for marketing.

we analyze the *term structure* of *interest rates* and discuss how the term structure is determined under certainty and uncertainty. We then discuss the concepts of *duration* and *convexity*. These concepts are basic to the notion of interest rate risk, so it is important to understand them before we discuss interest rate risk in detail, which we do next. Selected interest rate risk management techniques are subsequently examined. Next, we turn to liquidity risk, followed by concluding remarks. A case study is provided to illustrate some practical issues in interest rate risk management.

The Term Structure of Interest Rates

Review of Fixed-Income Valuation

What is the current value of a $250 riskless cash flow to be received in 1 year? We solve this problem by using the principle of *riskless arbitrage*. In particular, to prevent riskless arbitrage – which is essential in an efficient capital market – the price of this riskless cash flow in equilibrium must be related to the prices of other riskless instruments. In particular, suppose we observe that a United States government bond that promises $100 in 1 year is currently trading at $94.56. From this, we can deduce that the implicit 1-year return on riskless instrument is 5.75 percent (since $94.56 [1 + 0.0575] = $100). Thus, we should be currently willing to pay $250/[1.0575] = $236.41 for the riskless promise to receive $250 in 1 year.

But what if the riskless cash flow is promised to us 2 years from now? Well, then we have to find a riskless instrument of similar maturity (2 years) and payment characteristics (the only promised payment is 2 years from now and there are no interim payments). Suppose we observe that United States government "pure-discount" bonds with a 2-year maturity that promises a $100 payment are currently trading at $88.58. Then we can deduce that the 2-year riskless yield, on an annualized basis, is given by, i_o^2, where $\$100/[1 + i_o^2]^2 = \$88.58$. Solving this equation implies an annual two-period yield of $i_o^2 = 6.25$ percent. Thus, we get Figure 4.3.

That is, even though both the year 1 and year 2 cash flows are riskless, they have different discount rates applied to them. Why?

The reason is that future one-period interest rates are expected to *increase*. In our example, we know that the 1-year riskless rate at date 0 is 5.75 percent and the 2-year riskless rate at date 0 is 6.25 percent. We can infer the 1-year riskless interest rate, i_1^1, that is expected to prevail in the future at date 1. We can solve for it as follows:

$$\$221.45 = \frac{\$250}{[1.0575][1 + i_1^1]}$$

which yields $i_1^1 = 6.75$ percent. That is, the two-period rate 6.25 percent is the *geometric average* of the successive one-period rates, 5.75 percent and 6.75 percent.

Cash Flow	$250	$250
Discount Rate	5.75%	6.25%
Present Value at t = 0	$236.41	$221.45

FIGURE 4.3 Cash Flows and Discount Rates

The Yield Curve

What we have seen above is that interest rates on debt instruments of different maturities are related through investors' expectations about future interest rates. Our discussion deals with zero-coupon (pure-discount) bonds until we get to duration. A useful concept for this discussion is *yield to maturity* (YTM), which is defined as the internal rate of return that equates the present value of the future cash flows from a bond to the current market price of the bond. The relationships among the yields on different bonds are summarized by the *term structure of interest rates*. We define the term structure of interest rates (or the yield curve) as the relationship between the YTM and the length of time to maturity for *debt instruments of identical default risk characteristics*. It is critical to equalize the default risk of the bonds whose yields we are comparing. For simplicity, we will confine our attention to bonds without default risk. Thus, the *YTM* on a bond with m periods to maturity is defined as the annualized equivalent discount rate at which the cash flows from the bond must be discounted m periods to arrive at its market price. Figures 4.4 and 4.5 show two different yield curves, each describing the yields of bonds that are identical, except in maturity. The yield curve in Figure 4.4 is for U.S. Treasuries and is upward sloping. It is the "on the run" curve, in which the *implicit* zero-coupon yield curve is interpolated from full-coupon bond prices. The yield curve in Figure 4.5 is for German government securities. It is cup shaped. For shorter maturities, this yield curve is "inverted," that is, the YTM decreases with maturity. For intermediate maturities, it is virtually flat, that is, the YTM is almost independent of maturity in this range. And for longer maturities, the yield curve slopes upward, that is, the YTM rises with maturity.

What determines the shape of the yield curve? For simplicity, we will examine this question first in a world of perfect certainty. Uncertainty will be dealt with subsequently. In both cases we assume that a financial market equilibrium precludes *riskless arbitrage*.

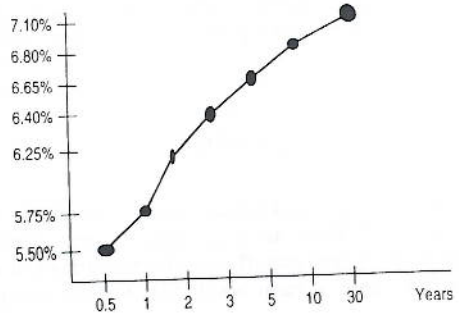


FIGURE 4.4 Risk-Free Term Structure for U.S. Treasury Securities as of July 25, 1996

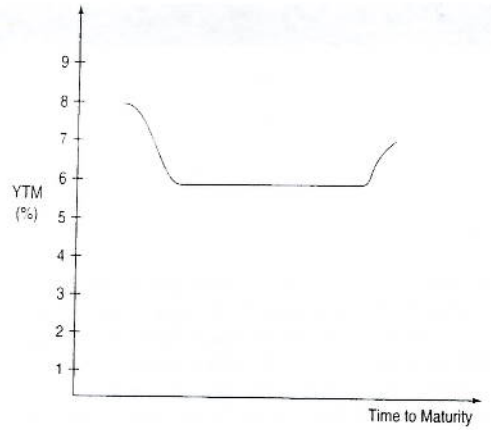


FIGURE 4.5 Yield Curve for Government Securities in Germany as of March 22, 1993

Yield Curve Determination Under Certainty

The Basic Model: Let P_t^m and i_t^m be the price and YTM, respectively, at time t of a bond of maturity m years. We assume the unit of time is 1 year, and all bonds are traded, so that prices are available from the market. As an illustration, we will examine the yield relationship between two bonds, one with a maturity of 1 year and the other with a maturity of 2 years. For simplicity, we will assume that each is a zero-coupon (pure-discount) bond and has a face value, F, of \$1. A zero-coupon bond makes a single promised payment (often called a balloon payment) at maturity, and no payments prior to that. Now, the YTM on the 1-year bond at the present time ($t = 0$), i_0^1, is the internal rate of return that discounts the \$1 face value over one period to equal the current market price of the bond.

$$P_0^1 = \frac{F}{1 + YTM} = \frac{1}{1 + i_0^1}. \tag{4.1}$$

Similarly, the YTM on the 2-year bond at $t = 0$, i_0^2, is the internal rate of return that discounts the \$1 face value over two periods to equal the current market price of the bond.

$$P_0^2 = \frac{F}{(1 + YTM)^2} = \frac{1}{(1 + i_0^2)^2}. \tag{4.2}$$

Now suppose we take \$1 today and invest it in the 2-year bond. Because it sells at $\$P_0^2$, we will be able to buy $1/P_0^2$ units of it. Then, 2 years from now (at $t = 2$), our investment will fetch us a (sure) payoff equal to the number of bonds we have bought $(1/P_0^2)$ times the face value of each bond (\$1). That is, our payoff at $t = 2$ will be [using (4.2)]

$$1/P_0^2 = (1 + i_0^2)^2. \tag{4.3}$$

Another use of our \$1 would be to invest it in the 1 year bond right now. We will be able to buy $1/P_0^1$ units of it at $t = 1$, then our payoff will be the number of bonds we have bought $(1/p_0^1)$ times the face value of each bond (\$1). That is, our payoff at $t = 1$ will be [using (4.1)]

$$1/P_0^2 = (1 + i_0^1). \tag{4.4}$$

What shall we do with this money at $t = 1$? Invest it, of course! Suppose we invest in another zero-coupon, \$1 face value, 1-year bond that will be issued a year from now (or equivalently, a multiyear bond with 1 year left to mature). Since we are currently in a world of certainty, we should be able to forecast the price, P_1^1, of this 1-year bond (issued 1 year from now) with perfect accuracy. With $\$(1 + i_0^1)$ to invest, we should be able to buy $(1 + i_0^1)/P_1^1$ units of this bond. Note that the YTM, i_1^1, of this bond is the internal rate of return that discounts the \$1 face value over one period to equal the current bond market price, and is thus

$$P_1^1 = 1/(1 + i_1^1). \tag{4.5}$$

Since we have bought $(1 + i_0^1)/P_1^1$ units of this bond at $t = 1$, and the face value of each unit is \$1, our payoff at $t = 2$ will be [using (4.5)]

$$[(1 + i_0^1)/P_1^1] \times 1 = (1 + i_0^1)(1 + i_1^1). \tag{4.6}$$

The Absence of Arbitrage and the Yield to Maturity Relationship: Equilibrium in this market requires that there be no riskless arbitrage opportunities. That is, we should not be able to do better at $t = 0$ with either the strategy of investing into the 2-year bond or investing in the 1-year bond and rolling over the proceeds into another 1-year bond. Both strategies should yield identical proceeds at $t = 2$ since we started out in each with identical \$1 investments. That is, the expressions in (4.3) and (4.6) should be equal. This gives

$$(1 + i_0^2)^2 = (1 + i_0^1)(1 + i_1^1),$$

or

$$(1 + i_0^2) = \sqrt{(1 + i_0^1)(1 + i_1^1)}. \tag{4.7}$$

Thus, the (annualized) YTM on the 2-year bond should be the *geometric average* of the YTMs on two successive bonds, each of maturity 1 year. This relationship is sometimes known as the *expectations hypothesis*, because it says that the yield on a long-term bond should be based on the expectations of investors about the yields on a sequence of short-term bonds. The general form of (4.7) for any arbitrary number of years, n, is

$$(1 + i_0^n) = \sqrt[n]{(1 + i_0^1)(1 + i_1^1)(1 + i_2^1)(1 + i_3^1)\dots(1 + i_{n-1}^1)} \tag{4.8}$$

Spot Rates and Forward Rates: The future yields, i_1^1, i_2^1, i_3^1, are known as *forward* rates, whereas the current yields, $i_0^1, i_0^2, \ldots, i_0^n$, are known as *spot* rates. Note that the forward rate for any period in the future can be defined with the help of a ratio of bond prices. To see this, solve (4.7) to obtain

$$i_1^1 = \frac{(1+i_0^2)^2}{(1+i_0^1)} - 1.$$

Now, substituting for $1 + i_0^1$ and $1 + i_0^2$ from (4.1) and (4.2) respectively, we get

$$i_1^1 = \frac{P_0^1}{P_0^2} - 1.$$

Similarly, we can obtain $i_2^1 = \frac{P_0^2}{P_0^3} - 1$, and so on. A one-period-hence forward rate can thus be thought of as the interest rate on a one-period loan starting at some future point in time. An n-period-hence forward rate is the interest rate on an n-period loan starting at some future point in time. The general formula for the YTM on a bond of maturity n periods to be issued t periods from now (that is, the n-periods hence forward rate for time t) is $i_t^n = \sqrt[n]{\frac{P_0^t}{P_0^{n+t}}} - 1$. We can see now how the shape of the yield curve is determined. If investors believe that short-term interest rates will keep rising, then $i_0^1 < i_1^1 < i_2^1 < \ldots < i_{n-1}^1$, so that $i_0^1 < i_0^2 < i_0^3 < \ldots < i_0^n$, and the yield curve will be upward sloping. On the other hand, if investors believe that short-term interest rates will keep falling, then the yield curve will be inverted, or downward sloping. Given a set of bond prices, we can compute the implied forward rates in the market as we do in the example below.

Notice that the geometric mean of 5 percent, 9.03809 percent, and 16.25469 percent equals the current 3-year yield of 10 percent. Likewise, the geometric mean

Example 4.1 Suppose there are three zero-coupon bonds that are identical in all respects except maturity. Each bond has a face value of $10 million. One of them matures a year from now and is currently selling at $9,523,809. The other matures 2 years from now and is currently selling at $8,734,386. The third matures 3 years from now and is currently selling at $7,513,148. Compute the YTM for each of the three bonds, plot the yield curve (assuming that you can interpolate smoothly), and compute the available forward rates.

Solution We will solve this problem in two steps. First, we will use the specified bond prices to compute the various date-zero YTMs. Second, we will calculate the implied forward rates for different maturities by computing ratios of bond prices.

Step 1 Using our previous analysis, we have

$9{,}523{,}809 = 10{,}000{,}000/(1 + i_0^1)$, which gives $i_0^1 = 0.05$ or 5 percent.

Similarly,

$8{,}734{,}386 = 10{,}000{,}000/(1 + i_0^2)^2$, which gives $i_0^2 = 0.07$ or 7 percent. And,

$7{,}513{,}148 = 10{,}000{,}000/(1 + i_0^3)^3$, which gives $i_0^3 = 0.10$ or 10 percent.

Step 2 We will now compute the implied forward rates. The data given to us are that $P_0^1 = \$9{,}523{,}809$, $P_0^2 = \$8{,}734{,}386$, and $P_0^3 = \$7{,}513{,}148$. Now,

$$i_1^1 = \frac{P_0^1}{P_0^2} - 1 = \frac{9{,}523{,}809}{8{,}734{,}836} - 1 = 9.03809\%,$$

and

$$i_2^1 = \frac{P_0^2}{P_0^3} - 1 = \frac{8{,}734{,}836}{7{,}513{,}148} - 1 = 16.25469\%.$$

of 5 percent and 9.03809 percent equals the current 2-year yield of 7 percent. In addition, the mean of the current 2-year yield of 7 percent and the 1-year rate 2 years hence of 16.25469 percent will equal the current 3 year-rate of 10 percent. Thus, all possible 3-year investment strategies should produce identical returns. Our analysis thus far has proceeded under the assumption of certainty. We now introduce uncertainty about future interest rates.

The Lure of Interest Rate Risk and Its Potential Impact

As we saw in our earlier examples, yields of bonds of different maturities can be different. In Figure 4.3 we depicted a case in which the 1-year yield is 5.75 percent and the 2-year yield is 6.25 percent. That is, if we buy the 1-year bond at date 0 and hold it until date 1, we get a return of 5.75 percent and if we buy the 2-year bond at date 0 and hold it until maturity at date 2, it will give us a return of 6.25 percent. The difference in returns, 6.25 percent−5.75 percent = 0.5 percent, is called the *term premium*. We may define an *m-period term premium* as the difference between the expected return on holding for a one period of a bond with maturity m + 1 periods at the time of purchase and the return on a bond of a one-period maturity. If term premiums are positive, then longer-term bonds should have higher expected returns. In a world of certainty, the term premium reflects simply investors' expectation that future interest rates will be higher than current rates. But in a world of uncertainty—in which interest rates fluctuate randomly—the term premium has two components:

one reflecting *expected* changes in future interest rates, and the other reflecting a premium demanded by risk-averse investors for bearing the risk (in holding longer maturity bonds) that future changes in interest rates will deviate from what is expected (this can be viewed as a premium for bearing interest rate risk).

The term premium is usually positive. This can be seen in Figure 4.6 below, which depicts the estimated 10-year term premium in the United States Treasury Bond market. This figure shows that term premiums have declined since 1990 and have fallen sharply since 2004. This suggests a greater willingness on the part of investors to hold longer maturity securities. Given investor risk aversion, this may be indicative of a lower perceived macroeconomic volatility.

Evidence of a positive term premium can also be seen in Table 4.1, which provides data on government bond yields in different countries.

The term premium is usually positive and creates a strong inducement for banks to mismatch their asset and liability maturity structures. By holding assets of longer maturities than their liabilities, banks can profit from a positive term premiums. This is the lure of interest rate risk. But this is risky too, as the following examples shows.

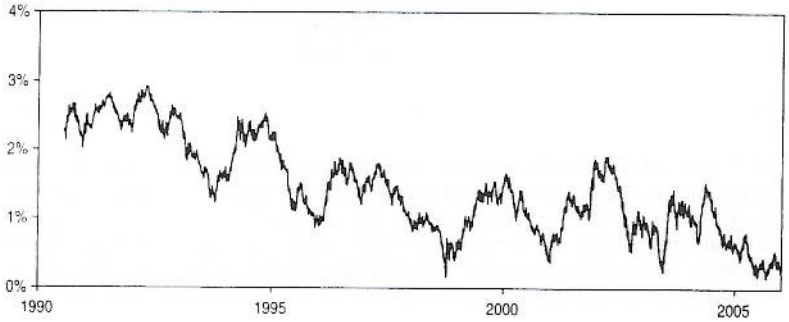


FIGURE 4.6 United States – Estimated Ten-Year Term Premium in the U.S. Treasury Market, 1990–2005

Note: Estimated instantaneous term premium at ten-year maturiy.

Sources: Don H. Kim and Jonathan H. Wright, "An Arbitrage-Factor Three-Factor Term Structure Model and the Recent Behavior of Long-Term Yields and Distant-Horizon Forward Rates," Federal Reserve Board, Finance and Economics Discussion Series Number 2005–33, August 2005; and the Federal Reserve.

TABLE 4.1 Government Bond Yields as of December, 2005

Country	2-year yield	10-year yield
United States	4.65%	4.8%
Euro Area	2.9%	3.5%
United Kingdom	4.42%	4.54%
Japan	0.20%	1.70%

Source: JP Morgan Economic Research, November 18, 2005.

Example 4.2 Suppose a bank's only asset is a 5 year United States government zero-coupon bond that promises to pay $100 million in 5 years. Its only liability is a 1-year $100 million certificate of deposit (CD). The yield to maturity (YTM) on 1-year riskless instruments is 5.75 percent and on 5-year riskless instruments is 6.65 percent.

This bank's balance sheet in economic value terms will look like this:

Economic Value Balance Sheet (in millions)

Assets		Liabilities and Equity	
Government bond	$72.48	CD	$70.92
		Equity	$1.56
Total	$72.48	Total	$72.48

The economic value of the government bond is $\$72.8 = \frac{\$100}{(1.10665)^5}$ whereas the economic value of the CD is $\$70.92 = \frac{\$100}{1.0575}$.

The economic value of the bank's equity is a plug and it arises from the term premium represented by the difference in the rates or return on the bank's assets and liabilities. As long as interest rates do not change, the bank will earn the term premium.

Now what happens to the value of the bank's equity if there is a parallel shift of the yield curve and all yields increase by 100 basis points? The new economic value balance sheet now looks like this:

Assets		Liabilities	
Government Bonds	$69.17	CD	$70.26
		Equity	−$1.09
Total	$69.17	Total	$69.17

The new economic value of the government bond is $\frac{\$100}{[1.0765]^5} = \$69.17$ and the new economic value of the CD is $\frac{\$100}{[1.0675]} = \$70.26$.

The equity value, which is a plug, is value of assets – value of liabilities = $69.17 −$70.26 = −$1.09.

So we see that even though there was only a modest and equal increase in all interest rates, the economic value of equity fell from $1.56 million to a negative $1.09 million. Why? The reason is that the long-term cash flow represented by the bank's asset has a value that is much more sensitive to interest rate changes than the short-term cash flow represented by the bank's liability. Thus, banks that are typically mismatched in a manner similar to our hypothetical bank – with assets of longer maturity than liabilities – experienced a decline in their equity values when interest rates rise. This kind of interest rate risk arises because a typical bank's assets and liabilities are mismatched in a particular way.

The existence of a positive term premium has profound implications for banks. On the one hand, it allows banks to profit from a maturity mismatch on their balance sheets. On the other hand, it imposes interest rate risk on banks. So, while the lure of profiting from maturity mismatching can be quite strong, the risk of mismatching can be ruinous, as many S&Ls and Orange County, CA, found out to their chagrin.

Could the bank have hedged its shareholders against interest rate risk by matching maturities? Not necessarily. The reason is that the banks need to match the *exact timing* of their asset and liability cash flows. Shorter-term cash flows behave differently than longer-term cash flows. To hedge its shareholders against interest rate risk, the bank must understand something about how asset and liability values will change, *given* changes in market yields. That is, the bank's shareholders will be protected against interest rate movements if, *for a given change in market yields,*

Percentage Price Change in Assets = Percentage Price Change in Liabilities
or

$$\left.\frac{\Delta P_A}{P_A}\right|_{\Delta_i} = \left.\frac{\Delta P_L}{P_L}\right|_{\Delta_i} \tag{4.9}$$

where ΔP_A = change in price of asset, P_A = price of asset, ΔP_L = change in price of liability, P_L = price of liability, and Δ_i = change in interest rate.

Let us now examine the value $\left.\frac{\Delta P_A}{P}\right|\Delta_i$.

Consider first a *flat* term structure, with $i = 10$ percent and a 10-year zero-coupon bond with \$100 par. How will the price of this bond change if yields (interest rates) change by 1 basic point?

$$P(\text{no change}) = \frac{\$100}{(1.10)^{10}} = \$38.5543$$

$$P|_{\Delta_i = +0.0001} = \frac{\$100}{(1.1001)^{10}} = \$38.5193$$

$$P|_{\Delta_i = -0.0001} = \frac{\$100}{(1.10009)^{10}} = \$38.5894$$

$$\left.\frac{\Delta P}{P}\right|_{\Delta_i = +0.0001} = -0.09\%$$

$$\left.\frac{\Delta P}{P}\right|_{\Delta_i = -0.0001} = 0.09\%$$

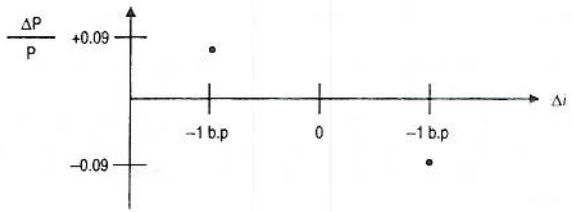


FIGURE 4.7 Price Changes for 1 Basis Point Change in Yields

Duration

The Inappropriateness of Maturity for Coupon-Paying Bonds

We saw that relative price change ($\Delta P/P$) is related to the yield change (ΔR). A mathematical relationship between $\Delta P/P$ and ΔR is given by *duration*, which is related to but different from maturity. The *maturity* of a bond tells the investor how long he must wait before receiving the terminal cash flow of the bond, or alternatively when the bond will mature or be redeemed. The maturity of a bond, however, does not give the investor all the needed information about the price volatility of the bond, unless it is a zero-coupon bond. This is because bonds of the same maturity can differ in their coupon payments through time. Moreover, in addition to coupon payments, bonds often provide other cash flows before maturity, such as amortizations. A bond that makes relatively large coupon payments early or amortizes rapidly has a shorter effective maturity than a bond that makes most of its large coupon payments late in the life of the bond. The reason is that the former generates much of its total cash flow well before its actual maturity date, whereas the latter skews its cash flows closer to its actual maturity date. We should, therefore, expect different sensitivities of the prices of these bonds to changes in interest rates. Note that we are now shifting our focus from zero-coupon bonds to bonds that may or may not pay coupons. All bonds we consider in our analysis are *nonamortizing*, that is, only coupon payments are received prior to maturity, and the entire principal is paid at maturity.

Duration Is the Answer

Duration, which is calibrated in the same temporal units as maturity, captures the timing of *all* cash flows generated by a bond, not just the terminal cash flow, and therefore is a more sophisticated measure of cash flow timing.[3] The duration of a bond is defined as the weighted average of the times to arrival of *all* scheduled future payments of a bond, where the weight attached to each payment reflects the relative contribution of that payment to the value of the bond. That is, each weighting factor is the present value of that payment divided by the present value of all payments of the bond. Consider a bond with N years to maturity, coupon payments $C_1, C_2, \ldots C_N$ where C_t is the coupon paid t years from now, and a principal (balloon) payment of B_N made at maturity. Let the term structure be *flat*, with *i* as the annual yield for all cash flows. Then the price of the bond at t = 0 is the present value of future payments:

$$P = \frac{C_1}{1+i} + \frac{C_2}{(1+i)^2} + \cdots + \frac{C_N + B_N}{(1+i)^N} \tag{4.10}$$

To see how P is related to R, let's take a derivative

$$\frac{dP}{di} = \frac{-C_1}{(1+i)^2} + \left[\frac{-2C_2}{(1+i)^3}\right] + \ldots + \left[\frac{-N[C_N + B_N]}{(1+i)^{N+1}}\right]$$

3. This concept was introduced by Macaulay (1938). Our treatment relies in part on generalizations by Fisher and Weil (1971), and Ingersoll, Skeleton, and Weil (1978).

or

$$dP = -\frac{-di}{1+i} + \frac{C_1}{(1+i)} + \frac{2C_2}{(1+i)^2} + \ldots + \frac{N(C_N + B_N)}{(1+i)^N}$$

Dividing both sides by P gives us:

$$\frac{dP}{P} = \frac{-di}{1+i}\left[\frac{\frac{C_1}{(1+i)} + \frac{2C_2}{(1+i)^2} + \ldots + \frac{N(C_N+B_N)}{(1+i)^N}}{\frac{C_1}{(1+i)} + \frac{C_2}{(1+i)^2} + \ldots + \frac{(C_N+B_N)}{(1+i)^N}}\right]$$

We can write this as:

$$\frac{dP}{P} = \frac{-di}{(1+i)}\left[1\left\{\frac{\frac{C_1}{(1+i)}}{\frac{C_1}{[1+i]} + \frac{C_2}{[1+i]^2} + \ldots + \frac{[C_N+B_N]}{[1+i]^N}}\right\} + 2\left\{\frac{\frac{C_2}{(1+i)^2}}{\frac{C_1}{[1+i]} + \frac{C_2}{[1+i]^2} + \ldots + \frac{[C_N+B_N]}{[1+i]^N}}\right\}\right.$$
$$\left. + \ldots + N\left\{\frac{\frac{C_N+B_N}{[1+i]^N}}{\frac{C_1}{[1+i]} + \frac{C_2}{[1+i]^2} + \ldots + \frac{[C_N+B_N]}{[1+i]^N}}\right\}\right] \qquad (4.11)$$

The numerator in each term represents a time of arrival, 1, 2, . . ., N, of a payment that is weighted by the present value of that payment. In the denominator, we have the present value of the sum of all cash flows promised by the bond, which should be its current market price, P. Define

$$W_t \equiv C_t/(1+i)^t \text{ for all } t = 1, 2, \ldots N-1 \qquad (4.12)$$

as the coefficient attached to the payment to be received t years from now.[4] Let $w_N \equiv (C_N + B_N)/(1+i)^N$. Then, using (4.12) and the definition of P, we can write (4.11) as

$$\frac{dP}{P} = -\frac{di}{[1+i]}\left[\frac{(w_1 + 2w_2 + 3w_3 + \ldots + Nw_N)}{P}\right] \qquad (4.13)$$

This equation gives the relationship between prices and yields. A fixed-income instrument's duration is its "price elasticity" and it relates percentage price changes to changes in yields. See Figure 4.8.

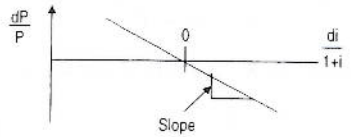


FIGURE 4.8 Duration

4. Each w_t is appropriately viewed as a "maturity coefficient" rather than a "weight" because the w_t's do not add up to one. However, each w_t divided by the denominator in (4.13) is a weight, that is, the $\hat{w}_t$s in (4.14) are weights.

Duration is the negative of the slope of the relationship shown in Figure 4.8. Thus, if we know the duration of an asset, we can predict its price sensitivity to a given change in yield. We can write:

$$\frac{dP}{P} = -D\left[\frac{di}{1+i}\right]$$

where D is duration. Defining $\hat{w}_t \equiv w_t/P$, we can write:

$$D = \sum_{t=1}^{N} t\hat{w}_t. \qquad (4.14)$$

Thus, (4.14) says that, to arrive at the bond's duration, we compute a weighted average of the times to arrival of its different promised payments, where the weight attached to each time to arrival is equal to the present value of the cash flow associated with that time to arrival divided by the price of the bond.

We can think of the duration of a bond then as a metric for the average number of years a holder of that bond must wait before recouping his investment. For risk assessment purposes, duration is a much more meaningful attribute of a bond than its maturity. The shorter the duration of a bond, the lower is its price volatility. Holding everything else (including the current value or price of the bond) fixed, an increase in coupon payments reduces duration, and an increase in maturity increases duration. A zero-coupon (pure discount) bond has the longest duration among bonds of the same maturity; indeed, its duration is equal to its maturity. These bonds have recently become very popular. One significant advantage that they offer is that all cash flows they generate (which are only maturity) are implicitly reinvested at the YTM, rather than at the prevailing interest rate as with coupon bonds. However, zero-coupon bonds are also very risky because of their longer duration and consequent higher price volatility. When interest rates are falling, the holder of a zero-coupon bond realizes a greater price appreciation than the holder of an otherwise similar coupon-paying bond. But when interest rates rise, the holder of the zero-coupon bond also experiences a greater price decline! Let us see the effect of duration at work in the following simple illustration.

Duration at Work: Some Numerical Examples

The following key points about duration are worth noting:

1. Duration is denominated in years. It is a measure of the "weighted average life" of the bond.
2. Longer maturity assets have longer durations, *ceteris paribus*.
3. For zero-coupon bonds, duration = maturity. For all other bonds, duration < maturity. Holding everything else fixed, an increase in the coupon decreases duration.
4. The duration of a floating-rate instrument ("floaters") where the coupon changes with interest rates is the time until the next repricing.

Example 4.3 Consider an interest rate environment in which the one-period annual yield is 10 percent and the two-period annual yield is 9.7824 percent, and suppose we have two riskless bonds (each with a 2-year maturity) that are identical in all respects except that one is a zero-coupon bond that matures 2 years from now and promises a balloon payment of $1,109.60, where the other is a bond that will pay a coupon of $100 1 year from now and another coupon of $100 plus a balloon payment of $900 2 years from now. Compute the durations of these two bonds.

Solution We solve this problem in three steps. First, we compute the current prices of the zero-coupon bond and the coupon-paying bond using the yield data provided. We find that both are equally priced. Second, we calculate the duration of the coupon-paying bond, which is less than that of the zero-coupon bond. Finally, in step 3 we compute the variances of possible price changes (due to random interest rate movements) and show that this variance is higher for the zero-coupon bond.

Step 1 The discount rate for one period cash flows is 10 percent and the discount rate for two-period cash flows is 9.7824 percent. Thus the price of the zero-coupon bond is

$$P_0 = 1109.6/(1.097824)^2 = \$920.64.$$

Similarly, the price of the coupon bond is

$$P_c = [[100/1.10] + [1000/(1.097824)^2]] = \$920.64.$$

Step 2 The above calculation shows that both bonds are equally priced. The duration of the zero-coupon bond is its maturity, which is 2 years. The duration of the coupon-paying bond is

$$D = \hat{w}_1 + 2\hat{w}_1$$

where $\hat{w}_1 = [100/1.10]/920.64 = 0.09875$

and $\hat{w}_2 = [1000/(1.097284)^2]920.64 = 0.90125.$

That is, 9.875 percent of the value of this bond is attributable to its first period coupon and 90.125 percent of its value is attributable to the sum of its second period coupon and principal. Hence, $D = 0.09875 + 2(0.90125) = 1.90125$ years.

5. The duration of a bank's "core deposits" is typically taken as zero.
6. The duration of a portfolio is the weighted average of the durations of all the assets in the portfolio.

Using Duration to Measure the Impact of Interest Rate Shocks on a Bank's Equity Value: Recall that a bank's balance sheet can be expressed as

$$A = L + E$$

Where A = assets, L = liabilities and E = equity. Then, given a change in yield Δi, the balance sheet changes can be expressed as:

$$\Delta A = \Delta L + \Delta E \quad [4.14]$$

Now:

$$\frac{\Delta A}{A} = -D_A\left[\frac{\Delta i}{1+i}\right]$$

which implies:

$$\Delta A = -D_A[A]\left[\frac{\Delta i}{1+i}\right] \quad [4.15]$$

Similarly,

$$\frac{\Delta L}{L} = -D_L\left[\frac{\Delta i}{1+i}\right]$$

which implies

$$\Delta L = -D_L[L]\left[\frac{\Delta i}{1+i}\right] \quad [4.16]$$

Assuming that the yield shock to the assets is identical to the yield shock to the liabilities, we can substitute (4.15) and (4.16) in (4.14) to obtain:

$$\Delta E = \left[-D_A[A]\frac{\Delta i}{1+i}\right] - \left[-D_L[L]\frac{\Delta i}{1+i}\right]$$

which implies:

$$\Delta E = \left[-D_A[A] + D_L[L]\frac{\Delta i}{1+i}\right]$$

or

$$\Delta E = -\left[D_A - D_L\left\{\frac{L}{A}\right\}\right][A]\left[\frac{\Delta i}{1+i}\right] \quad [4.17]$$

where ΔE is in dollars.

So, when market yields change, what drives the change in the bank's equity value? There are three main drivers:

1) The size of the shock $\left(\frac{\Delta i}{1+i}\right)$
2) The amount of the leverage the bank uses
3) The mismatch between the durations of the bank's assets and liabilities. The bank will be "immunized" when $D_A = D_L \cdot L_A$

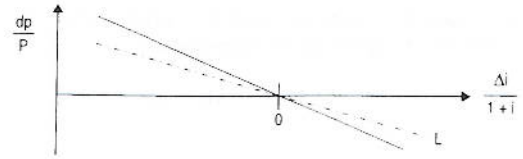


FIGURE 4.9 Asset and Liability Duration for Traditional Bank

How does this matter to a bank or a savings institution? To address this question, note that a traditional bank or savings institution has assets of longer duration than liabilities. Thus, its durations look like those shown in Figure 4.9.

What this means is that if yields increase, the bank's equity value declines (recall [4-17]), which shows that when $D_A > D_L$ and $L < A$, the term $\left[D_A - D_L \frac{L}{A}\right] > 0$, so $\Delta E < 0$ for any $\Delta i > 0$. If yields decrease, the bank's equity value increases. Thus, when a bank mismatches its balance sheet in the traditional way, it accepts interest rate risk in this way. Immunization closes the "gap."

A bank can alter its degree of immunization by changing the durations of its assets and liabilities. It can do this in two ways: on-balance sheet and off-balance sheet. On-balance sheet initiatives include making new types of loans, seeking new liabilities and changing its capital structure. Off-balance sheet initiatives include repurchase agreements, futures, options and swaps (we will discuss these in a later chapter).

Convexity

If a bank is interested in protecting its net worth against unexpected interest rate changes, duration matching can help; matching terms to maturity cannot do this unless all investments are of the zero-coupon variety. Suppose now that a bank is immunized and yields subsequently change. Does the bank remain immunized? The answer is no. The reason is that duration is an *approximation*. In fact, it is a *linear* approximation of a *nonlinear* relationship between prices and yields. We can see this with an example.

Example 4.4: Suppose we have a 10-year zero-coupon bond that is risk free, has a par value of $1,000, and is priced to yield 10 percent. What is its duration and how well will duration predict price changes if the yield moves up or down by 500 basis points?

Solution: Note that because this is a "zero" maturity = duration, so the duration here is 10 years. The current price of the bond is: $\frac{\$1{,}000}{(1.10)^{10}} = \$385.54$. Now consider the prices of this bond in response to a 500 basis point (b.p.) change in the yield.

Prices	Yield Change	
	+500 b.p.	−500 b.p.
Duration-Predicted Price: $\frac{\Delta P}{P} = -10\left[\frac{\pm 0.05}{1.05}\right] = \pm 47.62\%$	$\$385.54[1 - 0.4762] = \$201.95$	$\$385.54[1 - 0.4762] = \$569.13$
Actual Price	$\frac{\$1.000}{(1.15)^{10}} = \$247.18$	$\frac{\$1.000}{(1.05)^{10}} = \$613.19$
Error	−$45.23	−$44.78

We see then that duration *overpredicts* price declines when interest rates rise and *underpredicts* price increases when interest rates fall. Moreover, duration makes greater errors when yields rise than when they fall.

Why does duration make such prediction errors? The reason is that the true relationship between price changes and yield changes is *convex*, not linear.

When we first calculated the relationship between dP and di, we took a first derivative, which gave us the slope of the function in a "local" area, i.e., the slope of the curve, dP/di at di = 0. However, if we had gone further and computed the second derivative, we would have found $\frac{d^2P}{di^2} > 0$, i.e., all fixed-income securities are convex.

One implication of convexity is that duration will do a reasonable job in predicting price changes as long as interest rate changes are in the neighborhood of di = 0, i.e., relatively small changes like, say, 1 basis point. But the larger the interest rate change, the more erroneous duration is in predicting price changes. See Figure 4.10 below.

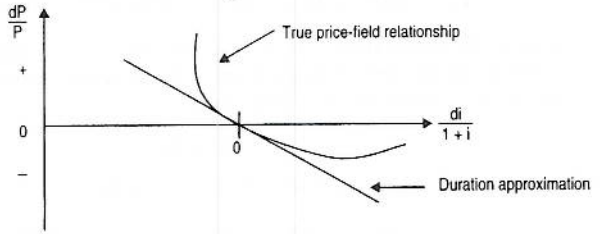


FIGURE 4.10 Price-Yield Relationship Is Convex

Implications of Convexity for Fixed-Income Securities and for Banks: There are three important implications of convexity for fixed-income securities:

1. The price decline given a rate increase is *smaller* than the price increase given a rate decrease of the same absolute magnitude as the rate increase.
2. Duration changes as yields change.
3. Greater convexity implies greater errors in the predictive ability of duration.

There are two important implications of convexity for banks:

1. Asset convexity is desirable. If the bank's asset portfolio is more than its liability portfolio, then properly done duration immunization never hurts the bank.
2. Duration immunization is a dynamic process since asset and liability durations change as yields change.

Interest Rate Risk

How Interest Rate Risk Can Affect a Financial Institution's Net Worth

The successful financial institution must understand its interest rate risk and manage the *durations* of its assets and liabilities. A *pure broker* need not worry about interest rate risk because its assets and liabilities are always duration matched. On the other hand, the *asset transformer* is often exposed to very subtle forms of interest rate risk. Consider the following simple example. A bank is borrowing and lending funds of two maturities: short term (1 year) and long term (2 years), all zero-coupon. Loans consist of $40 million short term and $40 million long term, while liabilities are $60 million short term and $10 million long term.[5] All numbers are in market value terms as of October 30, 2002. Hence, the bank's balance sheet is

Bank's Balance Sheet as of October 30, 2002

Short-term loans	$ 40,000,000	Short-term liabilities	$60,000,000
Long-term loans	$ 40,000,000	Long-term liabilities	$10,000,000
Total assets	$ 80,000,000	Total liabilities	$70,000,000
		Equity	$10,000,000
		Total equity and liabilities	$80,000,000

The yield curve as of October 30, 2002, is a flat solid line, as shown in Figure 4.11. Annual yields on assets and liabilities of all maturities are 10 percent.

Now suppose that on October 31, 2002, the yield curve shifts to the dotted line shown in Figure 4.11. All yields rise to 12 percent.

Each dollar of short-term assets (or liabilities) decreases in value to $0.9821428 and each dollar of long-term assets (or liabilities) decreases in value to $0.9646045. The new balance sheet in market value terms looks as follows

Balance Sheet as of October 31, 2002

Short-term loans	$39,285,712	Short-term liabilities	$58,928,568
Long-term loans	$38,584,180	Long-term liabilities	$ 9,646,046
Total assets	$77,869,892	Total liabilities	$68,574,613
		Equity	$ 9,295,279
		Total equity and liabilities	$77,869,892

Thus, the market value of equity falls by $704,721 or 7.047 percent. The shift in the term structure affects the values of *both* the assets and the liabilities, but it has unequal effects on assets and liabilities due to unequal maturity weighting or duration. To see this, note that the duration of short-term assets is 1 year and the duration of long-term assets is 2 years. The weights attached to the short-term and long-term assets are 0.5 and 0.5. Thus, the duration of the asset portfolio is $0.5 \times 1 + 0.5 \times 2 = 1.5$ years. Similarly, the duration of the short-term liability is 1 year and the weight attached to it is $60 million/$80 million = 0.75, while the duration of the long-term liability is 2 years and its weight is $10 million/$80 million = 0.125. Thus, the duration of the liability portfolio is $(0.75 \times 1 + 0.125 \times 2) = 1$ year.

5. You can easily verify that the asset and liability portfolios here have different durations.

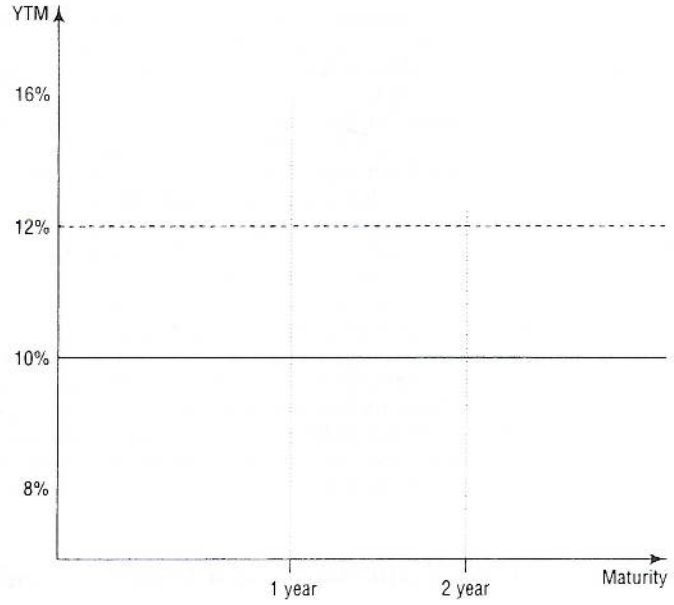


FIGURE 4.11 Yield Curves Facing Hypothetical Bank

While unequal duration weighting is risky, it is also a service provided by an asset transformer. By funding short (acquiring short-duration liabilities), the intermediary reduces the duration of its clientele's assets, thereby earning any term premium embedded in the yield curve. One simple way to eliminate interest rate risk altogether is to *equalize* the durations of assets and liabilities at all times. But then the institution forgoes duration/maturity transformation, a potentially profitable type of asset transformation.

A Case Study in Interest Rate Risk

Banks and other depository institutions often deliberately mismatch the durations of their asset and liability portfolios to profit either from term premiums or from their own expectations (guesses) about where interest rates are headed. Depository institutions characteristically fund their longer-lived assets with shorter-term liabilities. For instance, S&Ls historically funded 30-year fixed-rate mortgages with deposits

that were often subject to withdrawal on demand.[6] Similarly, commercial banks would finance 5- and 7-year fixed-rate "term" loans with demand and savings deposits, both of which could be withdrawn at a moment's notice. Such mismatches inevitably entail interest rate risk.

As an illustration, consider NCNB Corporation, a North Carolina-based banking company, which later went on to become Nations Bank and then merged with Bank of America; the bank speculated that there would be an interest rate downturn in 1990.[7] It thus lengthened the duration of its investment portfolio through 1989. At year end, the bank had a liability-sensitive balance sheet, largely because of its holdings of $6 billion in long-term Government National Mortgage Association (GNMA) mortgage-backed securities. As of December 31, 1989, about $1.5 billion more of NCNB's liabilities than its assets would have repriced over the next 12 months. If interest rates had fallen, NCNB would have enjoyed a huge profit. Instead, interest rates rose. As of year end 1989, the yield on 30-year GNMAs was 9.49 percent. By March 16, 1990, the 30-year GNMA yield was 9.95 percent. NCNB consequently suffered a $180 million unrealized loss in its bond portfolio. That news—plus disclosures in March 1990 that problem loans could rise by 25 percent in the first quarter of 1990—sent NCNB's stock plummeting from $46 in the first week of March 1990 to $40 by March 19, 1990, a decline of 12 percent.[8]

The Savings and Loan Experience and Other Episodes

Another striking example of the consequences of interest rate risk is the experience of the U.S. savings and loan (S&L) industry in the 1980s. S&Ls have traditionally financed themselves with short-maturity deposits and invested in relatively long-maturity, fixed-rate mortgages. Consequently, their liabilities repriced more frequently than their assets. As long as the yield curve sloped upward, this was a profitable maturity transformation. But in the late 1970s and early 1980s, the yield curve inverted as yields rose to historic highs. S&Ls took significant losses. This dissipation of much of the industry's net worth was the triggering event that led to the decimation of the industry years later. In particular, the loss of net worth meant that these institutions had much to gain and little to lose by pursuing risky investments. This led to further losses. The financial distress of Orange County in California in the 1990s is another example of the potentially devastating effect of interest rate risk.

Why Take On Interest Rate Risk?

The immediate question is: Why do banks and S&Ls *choose* to accept such exposure? That is, we have seen that it is possible for the bank to avoid taking much of the interest rate risk it normally takes on simply by matching the durations of its assets

6. Because of mortgage prepayments, 30-year fixed-rate mortgages have uncertain duration, typically of 7 to 12 years.

7. This discussion was reported by Kelly Holland in *American Banker*, March 20, 1990.

8. As Mr. John W. Munce, senior vice president and balance-sheet-management executive at NCNB put it, "We were postured to benefit from falling rates over a 12-month horizon. We definitely took some losses."

and liabilities, so interest rate risk is largely an avoidable risk. To answer this question, we go back to the theory of the term structure of interest rates. The presence of a risk premium in the term structure invites those who are more risk tolerant than the "average" (or representative) investor to hold long-term assets and fund these assets with shorter-term liabilities. Their reward is the premium in the yield curve that reflects the greater risk aversion of the respective investor.

Why should banks and other depository institutions be more risk tolerant than others? This is an issue we will take up in later chapters, but for now it suffices to note that deposit insurance may be one reason for a preference for risk on the bank's part. Of course, not all banks will desire to take on the same amount of risk. As in the case of their borrowers, the risk-taking propensities of banks depend on their own capital levels. Banks with more capital may wish to make investments that are less risky than those desired by banks with less capital.

The upshot of this discussion is not that an asset transformer should not take interest rate risk, but rather that such risk must be carefully assessed and managed.

Liquidity Risk

Our discussion of liquidity risk proceeds as follows. First, we introduce the concept of liquidity risk and discuss what liquidity risk means for a bank. We then present some formal definitions of liquidity. This follows with discussions of ways in which depository institutions manage liquidity risk. Finally, we end with a discussion of how a central-bank-based solution to the liquidity problems of individual depository institutions creates a moral hazard of its own.

What, After All, Is Liquidity Risk?

There are occasions on which the bank does not have ready access to funds that it needs, and is therefore forced to incur costs. These could be the costs associated with passing up investment opportunities. Alternatively, they could be distress financing costs. These are examples of situations in which the financial intermediary faces liquidity risk. We define *liquidity risk* as the risk of being unable to satisfy claims without impairment to its financial or reputational capital.[9]

Informational frictions are at the heart of liquidity problems. To see how informational asymmetries interact with default and interest rate risks to create liquidity risk, let us imagine that you own a bank that has made loans of $1 million with a maturity of 2 years and financed them with uninsured demand deposits. As a banker, you know more about the default risk of your loans than outsiders do, that is, there is asymmetric information about loan quality. Now, suppose that 6 months down the

9. It is important to distinguish between illiquidity and insolvency. The latter relates to a condition in which the value of the firm's liabilities exceeds the value of its assets, and hence its net worth is negative. Illiquidity can be as damaging and costly as insolvency, but it is a form of distress rooted in the (non)marketability of assets rather than in their ultimate or full value. To be sure, this may be a vacuous distinction when addressed at close range. Nevertheless, in thin markets, time and marketing efforts often are essential to the realization of asset values. Liquidating assets on short notice often results in "distress" prices. The relationship between time available for marketing and the realizable values of assets is central to the notion of liquidity.

road, $400,000 of deposits are withdrawn, but your existing stock of cash assets is only $100,000. This means you need to raise $300,000 to fund the deposit withdrawal. If potential depositors' perceptions about the quality of your loan portfolio are sufficiently favorable, you will not have any trouble acquitting new deposits in the amount of $300,000. But suppose that outsiders have received unfavorable information about your loans.[10] If this information is sufficiently unfavorable, new deposits may simply not be forthcoming,[11] or you might have to pay an excessively high interest rate—relative to the rate you consider "appropriate"—to attract the necessary deposits.[12] This is an example of liquidity risk.

There are two points we should note about this example. First, an informational asymmetry about asset quality plays a pivotal role in creating liquidity risk. If outsiders knew as much about your loan quality as you do, then you would be able to acquire the deposits you need at a price that you consider appropriate for the risk associated with the loan portfolio. This eliminates liquidity risk.

Second, duration mismatching may be an important ingredient in creating liquidity risk, but it is not a necessary ingredient. To see the importance of duration mismatching, suppose your asset and liability portfolios were perfectly duration matched. Then the assets that were funded by a specific set of liabilities would pay off at the same time that the liabilities came due, and informational asymmetry about these assets that arises *after* these assets are on the bank's books would not matter. Of course, if an informational asymmetry exists about the new loans you make, then a premium reflecting this asymmetry will show up in the interest rate on the deposits raised to fund these loans. However, you can pass this premium along to your borrowers in the way you price your loans, so that your capital is not impaired.

The Interaction Between Liquidity and Default Risks

However, you could have liquidity risk even with a duration-matched balance sheet. If some of the loans funded by deposits were to default, then withdrawals of these deposits would need to be funded in part by new deposits, assuming that loan defaults are large enough to leave insufficient liquidity to finance the withdrawals. Unless you plan to make new productive investments, depositors would have little reason to provide new deposits. Thus, new deposits would not be available *just* to finance old deposit withdrawals. To see this in the context of the previous example, suppose that both loans and deposits have a 2-year maturity. However, due to loan defaults, only $1 million is collected from loan repayments at maturity, whereas deposit withdrawals at maturity amount to $1.3 million. New deposits of $300,000 must be raised to finance withdrawals. This amount can only be raised against new assets that you acquire. Suppose now you wish to make $2 million in new loans with a 2-year maturity and thus need to raise $2.3 million in new deposits (ignore equity capital for now) that will also have a 2-year maturity. If your assessment of the quality (repayment probability) of these loans is higher than that of depositors in general, then the deposit interest rate will exceed what you believe is justified by the default risk of

10. This information may be different from what you know about your loans, that is, you may still know more than outsiders and may thus believe that your loan quality is good.

11. Indeed, it is possible that all of your existing deposits may be withdrawn.

12. In fact, your willingness to pay such a high rate of interest may be viewed as a signal of poor loan quality. Then, liquidity risk can be interpreted as the likelihood of incurring this signaling cost.

your loans. Suppose that, in present value terms, the excess amount you must pay in deposit interest is $46,000, that is, 2 percent of the total deposits acquired.

How much of this excess amount can you pass along to your borrowers? The answer depends on competition in the credit market. For simplicity, suppose that any other lender would face the same problem in communicating information about these loans to potential depositors, that is, any lender would suffer a cost equivalent 2 percent of the total deposits. However, a bank that does not need to finance *old* deposit withdrawals would need to raise only $2 million in deposits. Hence, 2 percent of $2 million can be passed along to borrowers in the form of a higher loan interest rate.[13]

Returning to your bank, then, if you are to be competitive in pricing your loan portfolio, you'll be able to pass along $40,000 in excess deposit interest to your borrowers. But that means you are stuck with a $6,000 "out of pocket" expense that arises because of your lack of sufficient liquidity to meet the excess of deposit withdrawals over net loan revenues. The possible incidence of such a cost is part of liquidity risk. Note again that this cost arises because there is a problem of asymmetric information about your loans. With perfect information, liquidity risk is not an issue here.

The Interaction Between Liquidity and Interest Rate Risks

We now turn to the interaction between interest rate risk and liquidity risk. There are two ways to explain this interaction. First, suppose we have deposit interest rate ceilings. Given this ceiling, a rise in market interest rates causes withdrawals because depositors can earn higher rates elsewhere. Hence, deposit interest rate ceilings transform interest rate risk into withdrawal risk.

Another way to understand this interaction is by returning to the example we discussed in the section under interest rate risk. If the term structure receives a random shock that causes interest rates to rise, it is possible that you will experience a deposit outflow as your depositors will want to reinvest their money at the prevailing higher interest rates. You have two ways to finance these withdrawals. One way is for you to acquire new (partially insured) deposits. But this may require you to pay a premium to depositors due to a possible informational asymmetry about your loan portfolio. Moreover, you must satisfy reserve and capital requirements on deposits. An alternative is to liquidate part of your asset portfolio to meet these unanticipated deposit withdrawals. You can do this by selling off marketable securities you hold or by selling off some of your loans.[14] Due to an informational asymmetry about your loans, however, you may only be able to sell your loans for less than what you think they are worth. The loss you incur as a result is also a part of liquidity risk. Although this loss is precipitated by an unfavorable move in interest rates, note again the

13. The assumption here is that there are many competing banks that can make the loans in question, and each of these banks needs to raise $2 million in deposits to finance $2 million in loans.

14. A bank can sell its loans to another bank just as a firm would sell its debt in a private placement. This practice, which is quite old, is known as "loan sales." A more recent practice is securitization, which involves the bank selling the loan, typically as a component in a portfolio of loans, directly to investors in the capital market. This is usually done through an underwriter and is a process of converting a previously untraded security into a traded security. We will have a lot more to say about this in Chapter 9.

central role played by asymmetric information. Moreover, the greater the asymmetric information, the greater the potential for loss, and hence the lower the asset's liquidity. This is why, despite an active secondary market, a corporation's common stock is not as liquid as a U.S. Treasury bill.

Some Formal Definitions of Liquidity

Think of P^* as the full-value price of an asset, that is, the highest price an owner can expect to realize by liquidating one unit, provided all useful preparations are made for the sale. If the asset is sold before all useful preparations can be made, a lesser price will be realized. Call this lesser price P_i, where $i = 0, \ldots, n$ indicates the time used for marketing, and n is the time needed to realize full value. The length of time used should be thought of as the interval between a decision to sell and the time at which a sales contract is consummated.[15] Hence

$$P_n = P^*$$

and for all values of $i < n$, the realized price of the asset, P_i, is less than full value. One way to think of liquidity is in terms of

$$L_1 = \frac{P_i}{P^*}.$$

A limitation of this definition is that the liquidity of a particular asset depends on the value of i chosen. Thus, for low values of i, one asset may be more liquid than another, whereas for greater values of i, the liquidity comparison might be reversed. This impedes the consistent ranking of assets according to their liquidity. One way to mitigate, if not obviate, this problem of liquidity reversal among assets is to think in terms of an "average" value of i. Hence

$$L_2 = \sum_{i=0}^{n} \frac{P_i}{P^*}.$$

A still more appealing approach recognizes the inherent uncertainty regarding i, the time interval between the decision to sell and the actual sale. Thus, we can view it as a random variable with a probability distribution, g(i), which stipulates the probability of each possible outcome ($i = 0, \ldots, n$). The expected value of an asset, E(P), is then defined as

$$E(P) = \sum_{i=0}^{n} g(i)P_i,$$

15. The terms of the transaction are fixed at the time the sales contract is consummated, but the transfer of property takes place at the "closing," a date that may coincide with the date of the sales contract, but often occurs later.

and this leads to a third definition of liquidity, which is

$$L_3 = \frac{E(P)}{P^*}.$$

The liquidity concept can be further generalized to account for marketing expenditures, say M. The more general view is that the realizable price of an asset depends on time, marketing expenditures, and full-value price, so that

$$P_i = f(i, M, P^*),$$

and if $\overline{M}$ is the optimally chosen marketing expenditure,

$$E(P') = \sum_{i=0}^{n} g(i)f(i, \overline{M}, P^*)$$

is the expected value of an asset, conditional on the owner's spending optimally on marketing. This leads to our fourth definition of liquidity

$$L_4 = \frac{E(P')}{P^*}.$$

and M/P^* can be thought of as a measure of the market's thinness, a measure akin to the bid-ask spread.[16]

Note that the positive relationship between available time for marketing and marketing effort on the one hand and realizable value on the other has nothing to do with changes in supply or demand for the asset; the realizable value increases in the context of given market conditions. Time is not used to await a more favorable market, but rather to do the marketing necessitated by costly information. For a depository institution, there are many ways to reduce liquidity risk. An obvious way is to simply keep more liquid assets on hand. The other is to reduce the deposit withdrawal risk that creates liquidity risk. A third way is to rely on a lender of last resort who stands ready to replenish the bank's liquidity when needed. In what follows, we discuss each in turn.

Reducing Liquidity Risk With Liquid Assets

Think of the fractional reserve banking system described in Chapter 3. That bank can be thought of as holding two kinds of assets: cash and loans that mature in two or more periods (prior to maturity the loans are assumed to be worthless). The bank's liabilities all mature in one period, and may or may not be renewed (withdrawn). If the fraction withdrawn after one period is equal to, or smaller than, the bank's holding of cash assets, the bank will continue in business for two periods, at least.

16. For a fuller development of this idea, see Greenbaum (1971).

On the other hand, if withdrawals exceed the bank's holding of cash assets, that bank will be unable to honor its liabilities—it has promised all depositors immediate access even though its own capacity to satisfy claims is strictly limited by its holding cash assets.[17] Therein lies the liquidity conundrum of banking.

Notice that an important role of a bank is the provision of liquidity services, and it provides this service by mismatching its balance sheet on the liquidity attribute, that is, it holds assets that are less liquid than its liabilities. This is one form of asset transformation. The *quality* of this liquidity service provided by the bank depends on three factors: the liquidity of its loan portfolio, the cash (or liquid assets) it has on hand, and the withdrawal risk in its deposit base. By investing in more liquid loans and/or keeping more cash on hand, the bank can improve its own liquidity. However, it does so at the expense of profits. An alternative would be to seek ways to dissipate withdrawal risk, which is what we turn to next.

Reducing Liquidity Risk by Dissipating Withdrawal Risk

A depository institution can reduce the variance of its deposit flows by diversifying the sources of funding, that is, having many distinct and dissimilar depositors. This is formally demonstrated in Appendix 4.2. A diverse depositor base results in more predictable deposit flows; the improved predictability reduces the cash needed to service a deposit base to any arbitrary probabilistic standard. That is, the larger and more diverse the depositor base, the smaller the cash holding necessary to achieve any preselected probability of a stock-out (liquidity crisis). This is one way the depository institution *produces* liquidity. Nevertheless, withdrawals will sometimes exceed the institution's capacity to service them, even though this may happen only with very small probability, and in that sense the system is imperfect. Indeed, this is the system's Achilles' heel. Bank runs are the trauma that illustrate this vulnerability of fractional reserve banking, a vulnerability caused by the illiquidity of bank assets.

Reducing the Liquidity Risk of an Individual Bank With a Lender of Last Resort

It was long ago discovered that the liquidity of a fractional reserve banking system can be ensured with a thoroughly credible "lender of last resort" (LLR). This was the major motivation for the creation of central banks, including the Federal Reserve System. With an institution capable of creating money limitlessly, it becomes possible to support banks facing the most extraordinary deposit outflows. Provided that the banks are sound (solvent, given reasonable time to liquidate their assets), this could

17. This is the rationale behind the standard measure of liquidity in the savings industry, which is the ratio of cash and short-term U.S. government securities and other specified securities to deposits and borrowing due within 1 year. The Office of Thrift Supervision (OTS) has established minimum liquidity requirements for savings institutions.

be done by having the central bank lend to the banks using their illiquid loans as collateral. With such a lending facility, sound but illiquid banks could be protected and financial market disruptions avoided. This argument is developed more fully in Appendix 4.2.

However, an inexpensive, readily available LLR faces the danger of inheriting the entire liquidity management problem of the banking industry. That is, the bank's incentive to hold cash assets (or even diversify its deposit base) is weakened if borrowings from the central bank are inexpensive and readily available. This is a moral hazard associated with the introduction of the LLR, and it has two implications. First, it shifts deposit seigniorage from the public to privately owned banks. Second, the LLR is also exposed to the credit risk of the bank's collateral. The moral hazard of lower, voluntarily held cash assets explains the consequent introduction of cash asset reserve requirements, and also why there are carefully administered detailed rules and informal restrictions governing access to the discount window.

Thus, legal reserve requirements and LLR pricing and availability shift at least a portion of the liquidity management problem back to the banks. Other banks, without access to an LLR facility, own the liquidity problem outright.

Closing Remarks on Liquidity

The management of liquidity is referred to as the treasury function, and it is usually entrusted to the chief financial officer (CFO). It is her responsibility to "fund the bank." This requires a professional understanding of the institution's cash flows, as well as all potential sources of liquidity. Ultimately, protection comes from maintaining diverse, capacious, and reliable sources of funding against future contingencies. This explains why the typical bank will borrow from virtually *all* reasonably priced sources. To be sure, cost will be a consideration, but opportunities to reduce short-run funding costs by concentrating on fewer funding sources are commonly avoided.

In "paying up" for funding diversity, the bank is purchasing lines of credit, and this reduces the likelihood of being rationed. It is common for funding sources to evaporate under stress; CFOs understand this only too well. Continental Illinois Bank and Trust found that holders of its large CDs (Certificates of Deposit) abandoned them in their hour of keenest need, and the high-yield bond market went into eclipse when Drexel Burnham Lambert was forced into insolvency because banks chose to withdraw their funding. The conventional protection against the trauma of being rationed is to accept the extra cost of participating in as many markets as possible, thereby diversifying funding sources. Liquidity is consciously purchased by banks as well as their borrowers, and it is the fragility of liquidity that makes this part of banking particularly challenging.

Conclusion

Like any other firm, a bank faces risks that can be managed but not totally avoided. For a bank, the three major risks are default risk, interest rate risk, and liquidity risk. These are interrelated and their interaction depends in an important way on

the presence of asymmetric information. The current approach is to manage these three risks and others holistically as part of Enterprise Risk Management (ERM).[18]

Interest rate risk is linked to the term structure of interest rates. Our analysis of the term structure both under certainty and uncertainty shows how yield and maturity are related. In both the certainty and uncertainty cases, the concept of riskless arbitrage plays a key role. Further, our analysis shows that the risk in holding a bond is more appropriately assessed in terms of its duration rather than its term to maturity. The definition of duration and the examination of its relevance in measuring the price volatility of bonds indicate how coupon-paying bonds should be analyzed. We also examined the concept of convexity and measures of interest rate risk exposure. We followed this with an examination of liquidity risk and the interaction of liquidity and interest rate risks. With these tools in hand, we considered the management of these risks by the bank.

Case Study Eggleston State Bank

Introduction

Mr. Edward Eggleston, CEO and primary stockholder of Eggleston State Bank, the bank he founded some 30 years ago in his hometown of Bloomington, OR, is worried. He has just gotten off the phone with an old friend of his, Fred Fisher. Fred had reported the difficulties he was having with his job search.

Fred's and Edward's life stories were remarkably similar. College roommates, they had both founded small hometown banks in the years following college and had managed to be quite successful for a number of years. But now, Fred is effectively wiped out—his bank has been closed by regulators and his fortune, invested entirely in the bank, has evaporated. Currently, he is going through the process of looking for a new job, maybe in the sort of big city he had always prided himself on avoiding.

Fred's bank had been fairly small, with $30 million in total assets, but had been consistently profitable as a small-town bank doing traditional banking—accepting deposits from individuals and small businesses in the short-term, while making long-term mortgage loans and business loans. But when state banking regulations were relaxed, allowing a branch of a major state bank to move into town, things got tighter. This competition, along with increasing volatility in interest rates and the bank's traditional mismatching of its balance sheet, led the bank into a situation with increasingly deteriorating capital, with a drop in capital over a 3-year period from $2 million to under $300,000. Finally, regulators moved in and took over the bank.

Edward Eggleston sighs, and wonders to himself whether the same thing could happen to his bank. His bank is much larger than Fred's with total assets of over $400 million (see Exhibit A). But with the rise of several regional banks with assets in billions of dollars, Edward is beginning to feel like he may face the same kinds of problems that beset Fred's bank, in the form of increased competition from larger, more sophisticated banks. He decides to meet with his executives to carefully

18. See, for example, Nocco and Stulz (2006).

investigate the exposure of Eggleston State Bank to interest rate risk, and to discuss the possibilities for hedging against changes in interest rates.

The Meeting

A week later, Edward Eggleston is sitting in his office with Carol Chipley and Douglas Date. Carol is a recent graduate of a top MBA program with strong analytical skills, hired in part to help modernize the bank's approach to risk management. Douglas, on the other hand, has risen to his current position from within the bank, primarily due to his sharp eye for detail and sound common sense.

Eggleston: O.K., gang. You both know our situation as well as I do. What I'm interested in is what options we have for action, and which you think we ought to pursue. Should we remain mismatched, or is it time for us to move into hedging?

Date: Well, as you know, Ed, I've always been skeptical about us getting involved in the latest fads in banking. After all, I don't see that we are so mismatched. Remember that article I showed you a while back, about a bank that started fooling around in the futures markets on the bad advice of a smooth-talking broker? I'm afraid that if we aren't careful, we could wind up making a big mistake. Besides, we've been here for 30 years now, steadily profitable. Why should we mess with a good system?

Chipley: I think that you are right, Douglas, when you say that we should be careful. But I think that for every story about banks losing money because a hedging program was poorly planned, we can find a dozen stories about banks that lost money, or even went under, because they weren't hedged at all. Plus, the banking environment has changed significantly in recent years. So what worked for the last 30 years might be fatal to us over the next 30 years.

Date: People are always saying that, but I don't really see what has changed. We've gotten bigger, but this is still a small-town bank. Our borrowers and our customers are mostly individuals and small to medium-sized businesses. Carol, weren't you just showing me the other day a chart showing how smooth our deposit flows have been over the past 5 years? (See Exhibit B). And the new administration seems committed to keeping Ft. Washington open, so it looks like the overall business outlook for the community is about the same as it ever was: stable and solid as a rock. This is a fairly prosperous area, after all. (See Exhibit C).

Chipley: Well, I'm not so sure that we can count on any administration keeping promises about military bases. But anyway, closing Fort Washington isn't the only risk that we face. I think that the increasingly competitive nature of banking means that world markets can affect what happens in our little town. Twenty years ago, our customers might not have worried so much about differences in interest rates; we were their hometown bank and we knew them and their business. But banking is more impersonal now, and we can't just expect our depositors to stay with us if we don't offer competitive interest rates. I think our investment and loan portfolios deserve a careful look (see Exhibits D and E).

Eggleston: Well, those are the reactions that I expected to hear from you. But I think that now is the time for some hard-boiled analysis. Let's sit down right now and come up with some likely interest rate scenarios. Then Carol can work with the figures and let us know exactly what would happen to the bank under a variety of circumstances (see Exhibit F).

The Numbers

Exhibit A
EGGLESTON STATE BANK
Year-End Balance Sheets (in Thousands)

	2004	2005
Assets		
Cash & due from banks	$59,696	78,645
U.S. govt. obligation	$38,612	45,284
Other govt. obligations	$58,030	49,456
Other securities	$6,678	6,439
Loans and discounts	$250,950	290,125
Bank premises	$12,698	21,924
Other assets	$2,996	2,876
Total assets	$429,660	494,749
Liabilities		
Demand deposits	$178,668	184,694
Time deposits	$122,164	166,995
Deposits of the U.S. govt.	$10,164	3,429
Other govt. deposits	$57,190	59,805
Due to commercial banks	$7,266	12,987
Total deposits	$375,452	427,910
Other liabilities	$23,520	34,925
Total liabilities	$398,972	462,835
Capital Accounts		
Common stock	$5,838	5,630
Capital surplus	$15,008	14,472
Undivided profits	$7,952	9,828
Reserves	$1,890	1,985
Total capital accounts	$30,688	31,915
Total liabilities and capital accounts	$429,660	$494,750

Exhibit B
Total Deposits (in Millions of Dollars)
(Expected Duration Six Months)

	High	Low	Daily Average
2001	305	257	284
2002	323	291	301
2003	363	323	357
2004	375	307	363
2005	427	375	400

Exhibit C
Market Area Economic Data
Income and Housing

Annual Household Income	Percentage of Households
Under $3,000	28%
$3,000–$6,999	20%
$7,000–$14,999	30%
$15,000–$24,999	21.5%
$25,000+	.5%
Home Ownership	
All Housing Units	30,000
Owner-occupied	51%
Rental	38%
Unoccupied	11%
Major Area Employers, Bloomington	
Ft. Washington	25,000
Lockheed	1,000
Kraft Foods	850
Bloomington College	730

Exhibit D
Eggleston State Bank
(Investment Portfolio, Today)

Description	Par Value	Coupon	Years to Maturity	Book Value	Bond Rating
U.S Government Securities					
Bills	2,500,000	—	8 months	2,235,000	—
Notes	4,000,000	6.00	2 years	3,765,000	—
Bonds	40,000,000	7.00	25 years	39,284,000	—
Other Government Securities					
Municipal Securities	50,000,000	6.00	22 years	49,456	Baa
Corporate Bonds					
Lockheed	7,000,000	12	17 Years	6,439	Aaa

Exhibit E
Eggleston State Bank
(Loan Portfolio, Summary Report, Today)

Borrower Type	Coupon	Estimated	Book Value
Short-Term Individual (Cars, and so on)	13.27	2.1	14,700,000
Short-Term Business	12.31	1.8	7,234,000
Medium-Term Business	11.45	5.3	42,300,000
Long-Term Business	10.4	7.9	78,766,000
Home Mortgages	8.3	9.1	179,000,000

During the meeting, the bankers came to an agreement on the following probabilities for the following scenarios:

Exhibit F
Likely Interest Rate Scenarios
(Scenario Names)

	Good	Bad	Ugly
Probability	.5	.3	.2
U.S. Govt. Securities			
Bills	11.00%	9.00%	12.00%
Notes	10.00%	10.00%	13.00%
Bonds	9.00%	11.00%	14.00%
Other Govt. Securities			
Municipal Securities	9.25%	11.75%	15.25%
Corporate Bonds			
Lockheed	9.75%	10.75%	13.75%
Loans			
Short-Term Individual	13.25%	11.25%	14.25%
Short-Term Business	12.25%	10.25%	13.25%
Medium-Term Business	10.50%	10.5%	13.75%
Long-Term Business	9.80%	10.75%	13.75%
Home Mortgages	9.00%	11.50%	14.50%

The Assignment

Eggleston: Carol, I'd like for you to take these numbers and report back to me on some very specific questions. What exactly is the extent of our mismatching? What would happen to the bank under the various scenarios that we've talked about? What kind of hedging program, if any, should we use to protect the bank?

Review Questions

1. What are the three major types of risks faced by banks?
2. What is the term structure of interest rates?
3. Under certainty, if the term structure is determined to preclude riskless arbitrage, what is the relationship between the yields on bonds of different maturities and why?
4. What is duration and why is it a more valid metric to consider for coupon-paying bonds than maturity? What is the relation between duration and price volatility for bonds with the same maturity?
5. What is convexity? Discuss its potential usefulness in evaluating bonds.
6. Discuss the pros and cons of duration mismatching for a depository institution.

7. What is liquidity risk and how is it linked to interest rate and credit risks? What is the role of asymmetric information in creating liquidity risk?
8. How can liquidity risk be managed? What are some of the impediments faced by banks in implementing an *integrated* risk management system that manages credit risk, liquidity risk, and interest rate risk?
9. Suppose there are three zero-coupon bonds, identical in all respects except maturity. Each bond has a face value of $1,000. One of them matures a year from now and is currently selling at $855.66. Another matures 2 years form now and is currently selling at $835.33. The third matures 3 years from now and is currently selling at $775.85. Compute the YTM for each of the three bonds, plot the yield curve (assuming that you can interpolate smoothly), and compute the available forward rates.
10. The annualized YTM on a single-period pure discount bond is 12 percent and that on a two-period pure discount bond is 10.45 percent. There are two bonds. One is a two-period, pure discount bond that promises a balloon payment of $1,200 at maturity. The other is a bond that will pay a coupon of $100 one period hence, and a coupon of $100 plus a balloon payment of $1,000 two periods hence. Compute the duration of these bonds and their possible price changes prior to maturity.
11. Given below is an excerpt from "A Friendly Conversation." Provide a critique.

Moderator: So, what do you people think? Will we ever really understand what happened to the American banking industry well enough to know what should be done?

Appleton: Well, I think banks and S&Ls were simply victims of the environment. We had an inverted yield curve—long rates were lower than short rates—for a while and this made it difficult for financial institutions to reap their normal profits from asset transformation; you know, I've never believed in the expectations hypothesis. It's a theoretical nicety with no practical relevance. Of course, the increased interest rate volatility didn't help. As if this wasn't enough, there was an enormous increase in competition, both domestic and international. These institutions must have felt like they were being squeezed by a powerful vise.

Moderator: By the way, Alex, I'll give you another reason not to like the expectations hypothesis—it's also wrong.

Appleton: I didn't know that. Are you sure? In any case, it's good to know you agree with me, Mike. But frankly, I'm surprised. Knowing how you and Beth feel about this, I thought I'd get more of an argument.

Moderator: Well, cheer up, Alex. My agreement with you is only partial. I agree that depository financial institutions faced a tough environment during the last 15 years or so. But I also think they could have *managed* their risks more intelligently. For example, they could have reduced the duration gaps in their asset and liability portfolios and made use of contemporary immunization techniques to hedge their interest rate risks. Like some of the investment banking houses, they could have been more innovative in brokerage activities, so that the resulting fee income would have made banks less dependent on the riskier asset transformation activities. Just look at the profits earned by some investment bankers who stripped Treasuries and sold zeros (pure discount bonds) like CATS (Certificates of Accrual of Treasury Securities) and TIGRS (Treasury Investment Growth Receipts). No, Alex! The real story

runs much deeper than your "passive victims of the environment" explanation. I think banks and S&Ls *exploited* the system and ripped off taxpayers.

Appendix 4.1 Dissipation of Withdrawal Risk Through Diversification

Suppose that a bank has n depositors, each of whom deposits \$1. Each deposit is subject to withdrawal after one period, but may remain for two. Assume that the probability that a \$1 deposit will be withdrawn after one period is one in ten, that is, $p = 0.1$, but whether a given deposit is actually withdrawn after one period cannot be known until that one period has passed.

Deposits are used to fund loans that pay back in full in two periods, but are worthless until they mature. (There is no secondary market in loans.) This is a harmless simplifying assumption and does not affect the argument that follows. Of course, the bank will need to hold some fraction of its assets in cash in order to satisfy its one-period withdrawals. The question is how much cash the bank should prudently hold. If the bank has \$1 or \$1 million of deposits, the probability of withdrawal remains fixed at 10 percent, and the expected withdrawal is this probability multiplied by the amount of deposits. However, if the bank has only \$1 in deposits, the withdrawal inevitably will be all or nothing at all, zero or one. Indeed, the expected value of \$0.10 is unattainable, and the bank's decision to hold 10 percent in cash, if feasible, is virtually pointless.

However, as the bank's depositors increase in number, assuming independence among them, the withdrawal of 10 percent becomes more predictable; in the limit, as depositors become more and more numerous, a 10 percent cash holding will "almost certainly" satisfy deposit withdrawals.

This idea is apparent from the definition of the standard deviation of a binomial distribution where n is the number of depositors and $q \equiv 1 - p$; the standard deviation of the bank's deposits will be $\sigma = \sqrt{npq}$.

Note that this measure of uncertainty varies with the square root of the *number of depositors*, and hence in the limit as the number of depositors increases to infinity, the standard deviation per dollar of deposit equals $\lim_{n \to \infty} (\sigma/n) = 0$.

This means that as the number of depositors becomes larger, the withdrawal uncertainty *per loan* diminishes, approaching zero in the limit, even though the withdrawal probability remains unchanged at $p = 0.1$. So, as the depositor population increases, the 10 percent withdrawal can be treated increasingly as a routine (almost fixed) cost, rather than as a potential catastrophe. The risk of ruin, the probability that withdrawals exceed the bank's cash holding, never actually becomes zero since $\sigma/n \to 0$ only in the limit. But the risk of ruin can be managed, and made indefinitely small by diversifying the bank's sources of funding.

Appendix 4.2 Lender-of-Last-Resort Moral Hazard

In a world of fiat money, value derives from an administered or artificial scarcity. That is, our money is money by fiat or legal mandate (hence legal tender) and is not convertible into gold or any other commodity at a fixed exchange rate, as in the case of commodity-backed money. The more money the government prints, or otherwise creates, the less its value, and this applies to bank deposits as well as to paper money. The administered scarcity of money also creates a monopoly profit referred to as "seigniorage." This profit on the production of money is shared by the privately owned banks and the public, via its effective ownership of the central bank. The Federal Reserve is nominally owned by member commercial banks. However, the equity in the Federal Reserve banks pays a statutorily fixed rate of return, much like a bond, whereas the residual earnings of the Federal Reserve flow back to the U.S. Treasury via a special franchise tax. Given that neither central bank nor private bank deposits pay interest (any interest rates below competitive rates will sustain the point), the distribution of seigniorage between the banks and the public (or central bank) depends on the cash asset reserves the banks choose to hold. The more reserves banks hold, the smaller will be banks' share of the seigniorage.

Since the introduction of an LLR reduces the amount of reserves the banks will desire to hold, it effectively shifts seigniorage from the public to the banks. This is the moral hazard associated with the introduction of an LLR, and it explains that one rationale for legal reserve requirements (that stipulate the minimum cash assets that banks must hold) is to restore the "appropriate" sharing of seigniorage between banks and the public.

This point is easily illustrated. Suppose we have a single commercial bank with \$10 million in deposit liabilities, an amount consistent with the money supply the central bank wishes to maintain in consideration of monetary policy. There are no reserve requirements and no LLR facility. The commercial bank voluntarily holds 10 percent of its assets in cash against withdrawal risk. It makes no difference whether the bank's cash assets are vault cash or deposits at the central bank, so for simplicity assume these assets are all on deposit at the Federal Reserve where they earn nothing. The commercial bank's balance sheet would then be

Commercial Bank

Cash assets	\$1 million	Deposit liability
Loans or other earning assets	\$9 million	\$10 million
Total assets	\$10 million	Total liabilities \$10 million

The Federal Reserve's balance sheet, to a first approximation, would show

Federal Reserve

Earning assets	\$1 million	Deposit liability \$1 million

Note that the Federal Reserve's deposit liability corresponds to the bank's cash assets. Now suppose the Federal Reserve introduces an LLR facility. It has no reason to change the money supply, but banks now have a new source of liquidity. Hence,

they will feel less need to hold nonearning cash assets. Say they cut these holdings from 10 to 5 percent. The bank's balance sheet now becomes

Commercial Bank

Cash assets	$0.5 million	Deposit liability	
Loans or other earning assets	$9.5 million	$10 million	
Total assets	$10 million	Total liabilities	$10 million

and the Federal Reserve shrinks to

Federal Reserve

Earning assets	$0.5 million	Deposit liabilities	$0.5 million

In effect, $0.5 million in earning assets have been transferred from the Federal Reserve's balance sheet to the bank's balance sheet, and this occurs as a direct consequence of the introduction of the LLR.

One could argue that if the LLR facility is properly priced, the moral hazard will be discouraged. However, note that before its introduction, the LLR interest rate was infinite, so that any finite interest rate will improve bank liquidity, and should therefore result in some reserve dissipation. As a historical matter, the LLR tends to price low for reasons that are not entirely clear. This generous pricing practice aggravates the moral hazard problem and heightens the need for legal reserve requirements.

Thus, reserve requirements control the moral hazard of the LLR, and a lowering of reserve requirements transfers deposit seigniorage from the public to the banks. Raising reserve requirements has the reverse effect. One hundred percent reserve requirements shift all deposit seigniorage to the public. This is the basis for the conventional wisdom that the reserve requirement is a tax on the banks, but one could just as easily argue that any reserve requirement less than 100 percent is a subsidy to banks. The hard question here is: To whom should the monopoly rents associated with administered money belong?

References

Cox, John, Jonathan Ingersoll, and Stephen Ross, "A Theory of the Term Structure of Interest Rates," *Econometrica* 53–2, March 1985, 385–407.

Fisher, Irving, and Roman L. Weil, "Coping with the Risk of Interest Rate Fluctuations," Journal of Business 44–4, January 1971, 408–431.

Greenbaum, Stuart I., "Liquidity and Reversibility," *Southern Economic Journal* 38–1, July 1971, 83–85.

Ho, Thomas S.Y., and Sang-bin Lee, "Term Structure Movements and Pricing Interest Rate Contingent Claims," *Journal of Finance* 41–5, December 1986, 1011–1030.

Holland, Kelley, "Capital: NCNB Loses Big Bet on Long-Term Rates," *American Banker*, March 20, 1990, 20.

Ingersoll, Jonathan E., *Theory of Financial Decision Making*, Rowman and Littlefield, New Jersey, 1987.

Ingersoll, Jonathan E., Jeffrey Skelton, and Roman L. Weil, "Duration Forty Years Later," *Journal of Financial and Quantitative Analysis* 13–4, November 1978, 627–650.

Klotz, Richard G., "Convexity of Fixed Income Securities," manuscript, updated.

Macaulay, F., *The Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856*, New York: National Bureau of Economic Research, 1938.

Nocco, Brian W., and René M. Stultz, "Enterprise Risk Management: Theory and Practice," working paper, Ohio State University, July 2006.

Sprenkle, Case, and Merton H. Miller, "The Precautionary Demand for Narrow and Broad Money," *Economica*, November 1980, 407–422.

CHAPTER • 5

Spot Lending

"Neither a borrower nor a lender be; For loan oft loses itself and a friend, and borrowing dulls the edge of husbandry."

William Shakespeare

Glossary of Terms

Loan: The extension of credit via a typically untraded and illiquid debt contract.

Security: A financial claim, debt, or equity, which may be traded or untraded.

COD: Cash on delivery as a method of payment for goods received.

Commercial Paper: Unsecured debt, offered as a short-maturity (less than 270 days) security by corporations.

T-bills, T-notes, and T-bonds: Debt securities of varying maturities issued by the U.S. government through the U.S. Treasury Department; hence, "T" for Treasury.

FHLB: Federal Home Loan Bank. The Federal Home Loan Bank System, headed by the Federal Home Loan Bank Board, was formerly the primary regulatory agency for savings and loan associations. The district home loan banks are now providers of financial services, including liquidity, to smaller commercial banks and thrifts.

FHLMC: This stands for the Federal Home Loan Mortgage Corporation. Also known as "Freddie Mac," its basic function is to facilitate the provision of liquidity to lenders by purchasing existing mortgages from their portfolios. It finances these purchases by borrowing from the Federal Home Loan Banks, issuing

GNMA-guaranteed mortgage-backed bonds, selling mortgage participation certificates on which it guarantees interest and principal, and selling guaranteed mortgage certificates.

FNMA: This stands for the Federal National Mortgage Association. It is a privately owned (stockholder-owned), government-sponsored enterprise. Also known as "Fannie Mae," its basic function is to provide a secondary market in trading and securitizing home mortgages. It is the largest purchaser of residential mortgages in the United States. Its activities are similar to those of Freddie Mac, except that it faces no statutory limitations on the organizations with which it can conduct business.

GNMA: This stands for the Government National Mortgage Association. This is a wholly owned, corporate instrumentality of the U.S. government, operating within the Department of Housing and Urban Development (HUD). Also referred to as "Ginnie Mae," its role is to enhance liquidity in the market for mortgages. Ginnie Mae does this in a variety of ways. For example, many mortgages carry a fixed interest rate so that when market interest rates rise, existing mortgages sell at a discount (that is, at less than face value). Ginnie Mae issues a commitment to the mortgage seller (the originating financial institution, for example) to purchase the mortgage at a fixed price. After acquiring the loan, Ginnie Mae sells it to "Fannie Mae" at the prevailing market price. Ginnie Mae absorbs any discount from the price paid to the seller. Another function of Ginnie Mae is to guarantee securities backed by government-insured or guaranteed mortgages. That is, Ginnie Mae provides guarantees for securitized claims against portfolios of government-insured mortgages.

S&P Stock Index: Standard & Poor's composite index of 500 large-company stocks.

Incentive Compatibility: A condition that requires the alignment of incentives between the agent and the principal. See Chapter 1.

C&I Loans: Commercial and industrial loans. These are loans extended to nonfinancial firms.

Nash Equilibrium: A steady state attained when none of the contracting parties has an incentive to change its actions unilaterally. See Chapter 1.

HLT: Highly leveraged transaction, which is a loan to a borrower with a very high debt/equity ratio.

Collateral: An asset used to secure a loan. Failure to repay the loan completely and in time transfers the collateral to the lender.

Absolute Priority Rule: A rule that prioritizes creditors' claims to a borrower's assets according to their seniorities.

GAAP: Generally Accepted Accounting Principles.

Prime Rate: A reference/benchmark borrowing rate posted by the bank for its better customers.

LIBOR: London Interbank Offer Rate. This is the interest rate banks charge each other for short-term loans in the United Kingdom.

CD Rate: The interest rate offered by banks on certificates of deposit.

Optimal Stopping Rule: A statistical decision rule that tells the decision-maker when to stop a sequential sampling process and make a decision. For example, a bank may have $1 million to lend and knows that the longer it waits, the more loan applicants it can screen before deciding who to lend the money to. However, waiting is costly because of the time value of money. An optimal stopping rule in this case would specify conditions under which the bank would find it most profitable to stop screening further loan applications. Another example is determining when a bank should stop acquiring additional information about a borrower, and make a decision.

Discriminant Analysis: A statistical technique used to identify the factors most useful in predicting an event. An example would be the factors useful in predicting bankruptcy.

The Glass-Steagall Act: An act passed by Congress in 1933 to separate commercial and investment banking in the United States. It prohibits commercial banks from engaging in securities underwriting and other investment banking activities as well as the activities of insurance companies.

Introduction

For many commercial bankers, lending is the heart of the business. Loans dominate asset holdings and account for a large share of revenues and costs. Lending takes place in both spot and forward credit markets. We begin here with a discussion of spot lending.

The purpose of this chapter is to explore the asset side of the bank's balance sheet. We begin in the next section with a brief review of the most prominent assets on a bank's balance sheet. The following section explains what we mean by lending, and the difference between loans and securities. We also discuss how these assets are purchased. The structure of loan agreements is discussed in a subsequent section. This is followed by a section that discusses the major informational problems in loan contracts and the importance of (perceived) loan performance for the determination of a bank's stock price. The next section examines credit analysis. Our emphasis is on the economic underpinnings of the various traditional factors considered in credit analysis. In particular, we relate these economic underpinnings to the informational problems pervasive in loan contracting. In the section that follows, we turn to sources of credit information. We consider both internal sources within the bank and external sources such as financial information agencies. In the next section, we take up analysis of borrower's financial statements. We follow it up with a section on the examination of loan covenants. Our focus is on the *why* of each covenant. A case study follows the concluding section.

Description of Bank Assets

Trends in the Composition of Bank Assets

There are three basic types of assets on a bank's balance sheet: loans, marketable securities, and cash. See Figure 5.1. Before we discuss each of these in detail, we will briefly review recent trends in the composition of bank asset portfolios.

In Figure 5.2 we show the time-series behavior of the composition of commercial bank assets. While loans have risen slightly as a fraction of total assets in the late 1970s

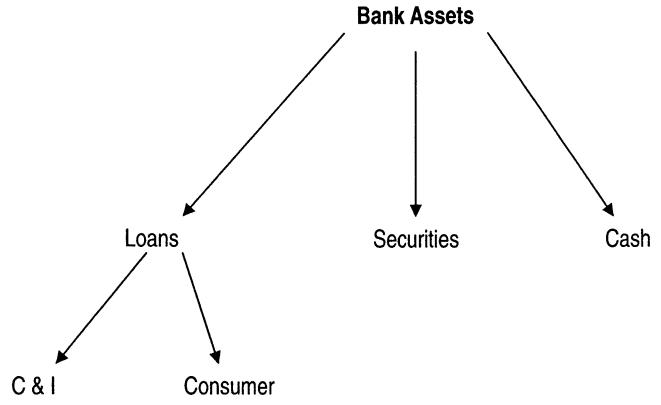


FIGURE 5.1 Spot Lending

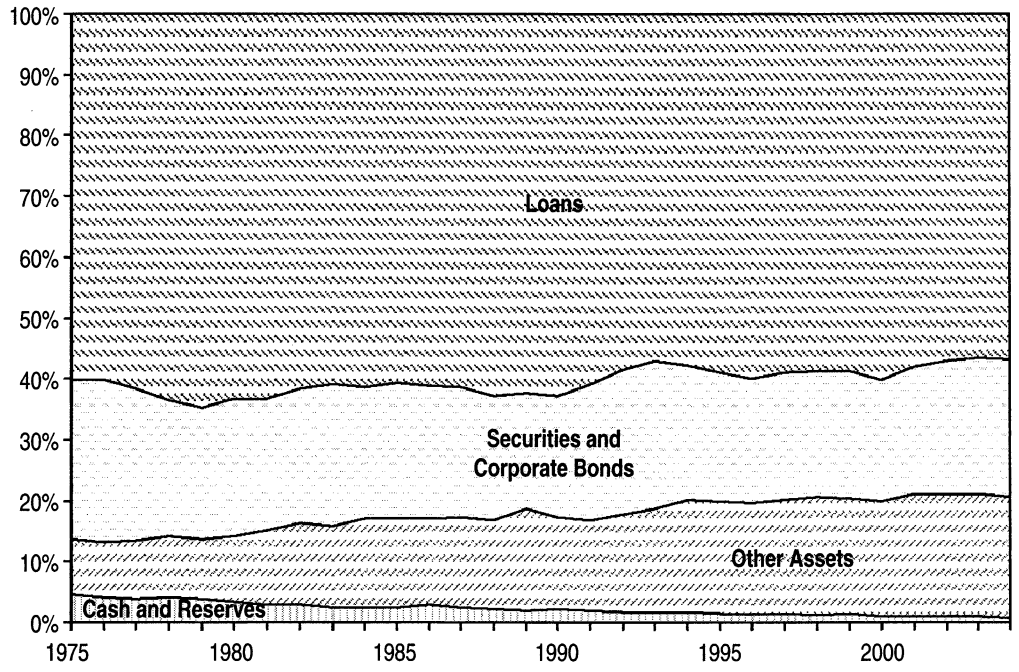


FIGURE 5.2 Composition of Commercial Bank Assets
Source: Federal Reserve Statistical Release: Flow of Funds Accounts of the U.S. 1975–1984, 1985–1994, and 1995–2004.

and 1980s, they declined slightly thereafter. Security holdings declined slightly in the late 1970s and have been relatively steady since. Cash and reserves have declined quite a bit, and this decline has been consistent through time. A clearer picture of what has been going on emerges from Figure 5.3, which shows the time-series behavior of commercial bank loans. It is apparent that C&I loans have declined in relative importance as banks have increased their mortgage holdings. Consumer credit has declined slightly in percentage terms from 20 percent in 1975 to 15 percent in 2004.[1]

1. Consumer loans are mainly comprised of credit cards, installment loans, mortgages, and home equity loans. These are essentially "commodity products," with apparently little product differentiation across banks. However, they still leave open considerable room for product innovation. For example, Wells Fargo gained prominence in the consumer loan market with its hybrid of a fixed-rate mortgage and an adjustable-rate loan. Moreover, the effectiveness with which credit information is processed is crucial in determining the attributes of consumers to whom these loans are made, and hence their profitability.

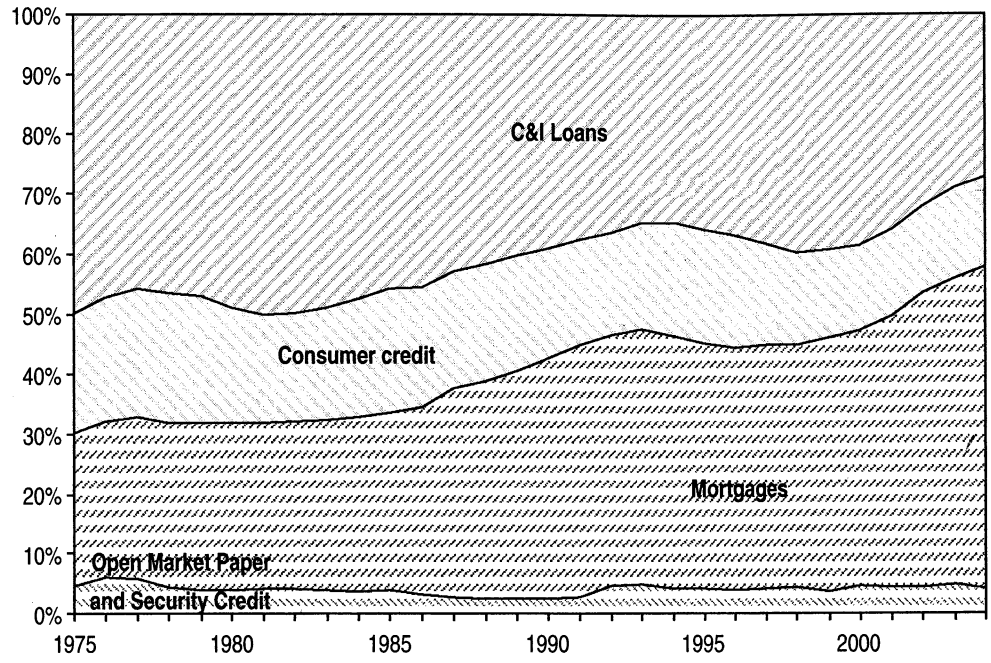


FIGURE 5.3 Composition of Commercial Bank Loans
Source: Federal Reserve Statistical Release: Flow of Funds Accounts of the U.S. 1975–1984, 1985–1994, and 1995–2004.

There are two main reasons for this trend. The first has to do with the changing nature of commercial lending. A bank has an advantage over the capital market in providing credit to a firm as long as banks have cheaper access to loanable funds than investors, *and/or* banks can resolve private information and moral hazard problems more effectively. Over the years, much of the deposit-related rents available to banks have eroded, thereby extinguishing virtually all of the funding advantage possessed by banks. Moreover, with the boom in financial innovation in the last two decades, a variety of new securities have been used by firms to raise funds directly from the capital market. These securities, as well as the securitization of bank-originated loans (see Chapter 9), have been designed to cope with the very problems of private information and moral hazard that banks have specialized in solving.[2] Thus, the relative advantage of banks over the capital market in providing credit to firms has diminished. With the capital market becoming a more viable source of competition in the commercial lending arena, the profitability of lending to large corporations has declined significantly for banks; hence, the relative decline in C&I lending.[3]

Types of Bank Loans

We will first discuss business loans, often referred to as C&I loans, which fall into four main categories.

2. For example, Green (1984) shows that a convertible bond (that is, a bond that can later be converted to stock by the bondholder) can be effective in controlling the moral hazard problem stemming from the borrowing firm's inclination to invest in risky projects to the detriment of bondholders.

3. For example, Security Pacific acquired $2.7 billion in mortgages in mid-1990 when it successfully bid for Gibraltar Savings, the largest California thrift under government control.

(a) **Transaction Loans:** A transaction loan is negotiated for a specific purchase and is tailored to the particular needs of the purchaser. The demand for these loans from a particular borrower is typically episodic and hence each loan is negotiated separately. The loan is usually secured by the asset being financed with the loan (for example, equity in another company), and repayment is expected to come from the use of this asset.

(b) **Working Capital Loans:** These loans are used by firms to finance routine day-to-day transactions. Thus, they are general purpose, short-term borrowings and are often used either to purchase current assets (like inventories) or to repay debts incurred in purchasing current assets. These loans are also usually secured by collateral such as accounts receivables or inventories.

(c) **Term Loans:** These are longer maturity loans used to buy fixed assets requiring large outlays of capital. Maturities typically run from 3 to 10 years. Repayment is normally amortized because it comes out of the cash flows generated by the asset financed with the loan. Borrowings are almost always drawn down under revolving lines of credit or similar commitments.

(d) **Combinations:** Working capital loans often include provisions that permit the conversion of short-term borrowings into term loans at the borrower's request.

We will now briefly review consumer loans.

(a) **Consumer Loans (excluding mortgage loans):** The most important types of consumer loans are direct loans and bank credit card receivables. A *direct consumer loan* is typically financing for the purchase of durable goods such as cars, boats, or appliances, and is secured with the asset being purchased. *Bank credit card* borrowings are a form of short-term, unsecured general purpose credit. Credit cards became widely used in the mid-1960s. Credit card lending has proved to be very profitable for banks.[4] The profitability of bank credit cards stems from three sources: (i) the discount at which the bank purchases sales slips from merchants (this discount typically ranges from 2 percent to 6 percent), (ii) the interest rate charged to a card user who chooses not to remain current in payments (most cards extend an interest-free grace period based on a monthly billing cycle), and (iii) the annual membership fees charged to credit card users.[5]

(b) **Mortgage Loans:** These are a specialized form of consumer and commercial lending. The purpose of a mortgage loan is to finance the acquisition or improvement of real estate. These loans are almost always secured by the real estate they finance. The three principal types of mortgage loans are: residential mortgage loans, construction loans, and commercial mortgage loans.

Until the advent of securitization, mortgage loans were illiquid assets because of the uniqueness of each property, the severity of private information problems, and the uncertain maturity of the loan due to the possibility of prepayment by the borrower. However, securitization took care of many of these impediments to the marketability of mortgages and facilitated the liquification of these instruments. This

4. See Ausubel (1990).
5. Many banks waive these annual fees because of increased competition for credit card business.

was especially true in the market for residential mortgages where Fannie Mae and Freddie Mac led the way under government auspices.

Securitization is a technology for transforming illiquid loans into traded liquid securities by separating the origination of the instrument from its funding. Typically, a financial institution such as a bank "originates" the loan, that is, it screens the applicant, designs the loan contract, and determines the pricing parameters. However, instead of using deposits to fund the loan as in the traditional case, the bank sells the loan to a special trust that assembles a portfolio of loans and funds the portfolio in the capital market, often with the advice and assistance of an investment banker. The services provided by the investment banker include the sale of claims against the loan portfolio to investors and then the maintenance of a secondary market in the securitized claims. The enormous growth in securitization in the past two to three decades is evidence of its benefits in the mortgage market. These benefits stemmed from the liquidity created by the standardization, diversification, possible subsidies provided by the government via Fannie Mae and Freddie Mac, and new contract design that accompanied securitization. Securitization is discussed in greater detail in Chapter 9.

Earlier, fixed-rate mortgages—in which the borrower's interest rate is fixed over the life of the mortgage—dominated the market. However, since the legalization of adjustable-rate mortgages (ARMs) in the 1980s there has been an explosion in the variety of mortgage designs. The terms of mortgages are as varied as the needs of borrowers and the imagination of lenders.

Marketable Securities Held by Banks

(a) Bankers Acceptances: These instruments arise mostly in connection with international trade. A bankers acceptance is a bank-guaranteed indebtedness of the bank's customer to a third party. This instrument usually arises as a time draft written by a firm in order to pay for some goods either in local currency or in foreign exchange. The draft is then "accepted" by the bank, that is, the bank guarantees its face value at maturity. The acceptance is then either held by the bank or sold in the secondary market and may be held by another bank. The originating bank typically charges a fee for the guarantee (acceptance) that is independent of the interest paid on the borrowing. Maturity is usually less than 6 months.

A bankers acceptance facilitates trade between parties that operate in different legal systems with wide geographical and cultural separation. If the exporter does not know the importer well enough, it will not ship goods, even on a COD basis. However, it is likely that the importer's bank is better known and hence its willingness to guarantee payment—which serves the purpose of substituting its own credit risk for that of the importer—facilitates trade. The bank issuing the guarantee also can be expected to know more about the importer, usually a customer of the bank. Its informational advantage vis-á-vis the exporter allows the bank to earn a fee on the acceptance. Thus, bankers acceptances are closely tied to the bank's role in providing a more efficient resolution of informational problems. For more on this, see Chapter 8.

(b) Commercial Paper: This is unsecured debt issued on the strength of the issuer's name. It is sold on a discounted basis like Treasury bills,[6] with maturities ranging

6. There is no explicitly stated interest rate, but the claim is sold at a price less than its face value (value at maturity), the difference implicitly defining the interest cost. Note, however, that discount yields are not directly comparable to bond yields; a translation is required to achieve comparability.

from 3 to 270 days and interest rates typically lower than prime and comparable to those on CDs and bankers acceptances. Only the best-known firms issue commercial paper because it is sold *directly* to investors, without an intermediary to resolve informational problems.

(c) U.S. Government Securities: These are important instruments for commercial banks because of their default-free nature and the highly liquid markets in which they are traded. As we saw in Chapter 3, private information content undermines liquidity, so U.S. government securities—which embody virtually no private information—provide banks with liquidity.

Income from all U.S. government securities is subject to federal income taxes as well as capital gains tax, but is exempt from state and local income taxes. Marketable U.S. government securities are of three types: bills, notes, and bonds. Treasury bills (T-bills) are short-term U.S. government securities (with original maturities of 91 days, 182 days, and 1 year) that, like commercial paper, are sold on a discounted basis. Treasury notes are similar to T-bills except that they have maturities not less than 1 year and not more than 7 years. Treasury bonds are issued with original maturities that often exceed 10 years, and can be as long as 30 years.

(d) U.S. Government Agency Securities: These are certificates of indebtedness issued by agencies of the U.S. government, such as the Federal Intermediate Credit Bank, the Federal National Mortgage Association (FNMA or Fannie Mae), the Federal Home Loan Bank (FHLB), and the Government National Mortgage Association (GNMA or Ginnie Mae). They are not direct obligations of the U.S. government, and they typically trade at a small premium over Treasury debt. Income on these securities, like direct U.S. government obligations, is exempt from state and local taxes, but not from federal taxes.

(e) State and Local Securities and Municipal Bonds: These debt instruments usually have a higher after-tax yield than Treasury and agency securities of comparable duration because of higher default risk and weaker liquidity. Their interest payments are exempt from federal income taxes as well as from home-state and local taxes. State and local government bonds can be divided into three broad categories: housing authority bonds, general obligation bonds, and revenue bonds. Housing authority bonds are issued by local housing agencies to build and administer housing. They are guaranteed by the federal government and are therefore virtually riskless. A bond is called a general obligation bond if the full faith and credit of the issuer stands behind the debt. In contrast, the interest and principal of a revenue bond is supported solely by the cash flow of a designated public project or undertaking. The revenues supporting these bonds may come from: (i) specifically dedicated taxes such as those on cigarettes, gasoline, and beer, (ii) tolls for roads, bridges, and airports, (iii) rent payments on buildings, office spaces, and the like. Typically, the bond payments are linked to the revenues produced by the project the bonds were used to finance.

(f) Other Assets: These include vault cash and deposits at the Federal Reserve, equity in subsidiaries, physical capital like buildings, computers, and loans originated by other banks that may have been acquired by the bank as part of a loan sale or through securitization. For short periods of time, the bank may also possess a variety of other assets acquired as collateral from delinquent borrowers.

What Is Lending?

A *Definition*

What is a *bank loan*? Simply put, it is the purchase of an asset (the borrower's indebtedness) that is typically an illiquid and highly customized financial claim against the borrower's future cash flows. In effect, the bank is obtaining from the borrower the legal right to a prespecified portion of the borrower's future cash flows over a prespecified period of time, and paying the borrower the present value of these cash flows. The bank's claim represents the borrower's repayment obligation and the loan amount represents the present value of these future obligations, assuming no extraordinary profit for the bank.

Methods of Acquiring Loans

There are two principal methods by which banks acquire loans: through *spot market purchases* and through *forward market purchases*. In the spot market, the bank can either originate the loan and then fund it by keeping the loan on its own books, or it can purchase the loan from another intermediary that originated it. A spot loan is created when the bank extends credit to a loan applicant immediately upon approval of the application. In the forward market, the bank issues a *promise* to the applicant that it will lend in the future on prespecified terms. Such a promise is known as a *loan commitment*. The bank commits to lend to the borrower up to a certain amount in the future on terms that are prespecified and at the option of the borrower. In this case, the bank is committing to purchase a financial claim from a particular borrower at some time in the future.

We discuss these two methods of asset acquisition in separate chapters. Spot market purchases and forward lending are covered in separate subsequent chapters. This division is merely for expositional convenience. In practice, the volume of spot and forward lending are inextricably linked. The extent of spot lending by the bank depends on how many of its outstanding loan commitments sold in previous periods are exercised or taken down in the current period. In general, a higher volume of takedowns of outstanding loan commitments implies a lower volume of spot lending in the current period, although the *total* volume of lending in the current period may rise (relative to that in the previous period) because of an unexpectedly high take-down on previously made commitments. This follows from the size constraints on banks associated with financial and human capital limitations.

The Decomposition of the Lending Function

The Decomposition: The subtlety of lending transactions is often blurred in the bundling together of distinct services relating to credit transactions. The normal commercial bank loan is logically decomposable into origination (the broker), funding (the lender), servicing (the collector), and risk-processing services (the guarantor). And lending can be thought of largely as credit risk management that includes these four activities as well as the bank's credit culture. See Figure 5.4.

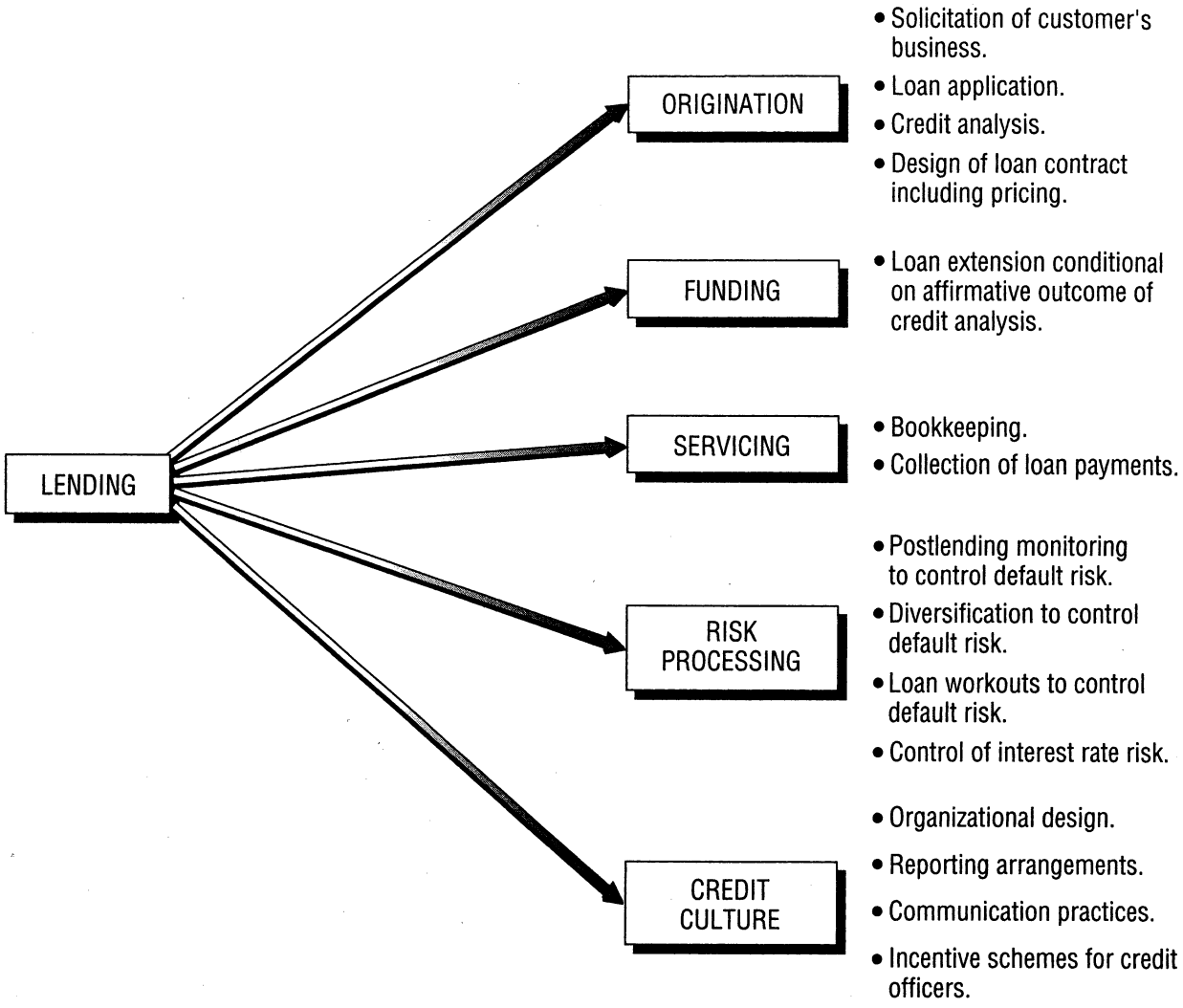


FIGURE 5.4 Decomposition of the Lending Function

Origination involves the activity of initiating a loan to a borrower. It is often described as the initial solicitation of the borrower and the screening of the loan application by the bank. Origination includes credit analysis and the design of the loan contract, both of which we discuss at length later in this chapter. *Funding* is the actual extension of the loan after an affirmative decision is reached in the credit analysis process. *Servicing* involves collecting loan repayments and keeping records. *Risk processing* involves postlending monitoring to control default risk, as well as activities designed to control the bank's interest rate risk arising from a loan duration that differs from the duration of the bank's liabilities. The *credit culture* involves the bank's organizational design, reporting arrangements, communication practices, and incentive schemes for credit officers. We will discuss credit culture later in the book. Much of our focus in this chapter and the next will be on origination (in particular, loan contract design and credit analysis) and risk processing (in particular, the control of default risk).

Industry Specialization: In the thrift industry (savings and loan associations, mutual savings banks, and the like) different institutions provide distinct credit services, which is clear evidence of institution specialization. For example, the mortgage banker originates loans and the mortgage processor services the loans. The loan is typically funded by the public (the net saver or surplus spending unit) in the form of newly purchased savings and loan deposits or mortgage-backed securities. The bulk of the credit and interest rate risk is sustained by savings and loan stockholders, the U.S. government (FDIC, FSLIC, NCUA, GNMA), specialized private insurers (for example, Mortgage Guarantee Insurance Corporation), or some combination of the three (FNMA, FHLMC). In commercial banking, it is common for the bank to hold originated loans. Consequently, the origination, servicing, and risk absorption is evidenced by an earning asset on the bank's balance sheet. The bank depositor holding a risk-free asset is funding the bank loans. The government, through the FDIC and the bank stockholders, shares the risks (uninsured depositors may sustain some exposure as well). Should the bank sell a loan, say to a closed-end mutual fund (as in the case of a savings and loan association selling mortgages to FHLMC or Salomon Brothers for packaging into a mortgage-backed security), then the security holder would do the funding and the location of the risk would depend on the specific terms (recourse or nonrecourse) of the sale. Irrespective of the terms of the sale, however, the bank need show no earning asset on its balance sheet and virtually all the same services would have been performed and the same exposures sustained without any accounting evidence thereof. This statement requires some qualification in that if a loan is sold with recourse, the accountant will probably insist on booking the asset, but if a loan is sold without recourse and a letter of credit is issued insuring against default (the above are equivalent), the balance sheet will show no loan and the letter of credit will probably appear in a footnote to the balance sheet, but not in its body.

In fact, banking reserve and deposit insurance premiums provide banks with an incentive to sell, rather than hold, earning assets. In this way, the bank can avoid these costs.

The traditional subsidy inherent in deposits (owing to underpriced deposit insurance, Regulation Q, and entry restrictions) encouraged banks to *hold* earning assets whereas deposit insurance premiums, reserve and capital requirements, along with less explicit regulatory costs, were a partial offset to the deposit subsidy. However, the deposit subsidy is rapidly disappearing, whereas many of the regulatory costs remain. Thus, we can predict that banks will de-emphasize the holding of the loans they originate, service, and guarantee. The recent emphasis on "fee income" is a reflection of this phenomenon.

Loans Versus Securities

In the previous discussions, we have talked about *loans* and *securities* as two distinct claims. The way we have defined loans, there is little difference between loans and debt securities, except that the latter are usually more liquid. That is, securities are traded in secondary markets, whereas loans usually are not. Loans are essentially *private* debt placements with banks. You will recall from our discussions in Chapter 4

that liquidity and marketability are interrelated. From an economic viewpoint, the distinction between loans and securities is in their *relative liquidity*.[7]

Viewed in this light, recent developments in the loan market can be seen as narrowing the distinction between loans and securities. We refer to *loan sales* and *securitization*. A loan sale, which is a fairly old practice, is simply the selling of a loan by the originating bank to an alternative funding agent, usually another bank. This can either be an *outright sale* of the loan, where the loan may have been originated by a single bank or as part of a *loan syndication*. With an outright sale, the originating bank disengages itself from the loan, that is, it makes the initial loan and then turns around and sells it to another bank, thereby removing the loan from its own balance sheet. A fee is earned for the originating service, so that the transaction leaves its mark on the originator's income statement. With an outright loan sale, the bank acts as a pure broker, although in practice almost every loan sale involves the originating bank retaining a part of the loan, so the bank is not a pure broker. Some loans are also made under syndication arrangements in which case there is *joint* origination of the loans by several banks. These loans may then be sold to others. Again, the "lead banks" in the syndicate earn fees.

Thus, loan sales enhance loan liquidity, especially if the originator maintains a secondary market. This blurs the distinction between loans and securities. A more recent practice for improving loan liquidity is securitization, which we discuss in detail in Chapter 9. Both loan sales and securitization trivialize the distinctions between loans and securities.

Structure of Loan Agreements

Trends in Loan Agreements: Commercial bank lending was once a fairly simple business. Most business loans were short-term, self-liquidating working capital credits, and terms were often left to informal agreements between a bank and its customers. Business lending began getting more complex in the 1930s when banks started making loans with maturities of more than a year, so-called term loans. Relations between banks and business borrowers have been growing more complex—and more formal—ever since.

Part of the push for more formality and variety in the design of agreements comes from the need for banks and borrowers to protect themselves from movements in interest rates over the credit cycle. Increases in market interest rates boost the costs to banks of funding outstanding loans and also reduce the attractiveness of existing credits. Reductions in market interest rates, on the other hand, often trigger prepayments.

7. From a legal standpoint, however, the distinction between a loan and a security was crucial during the Glass-Steagall Act, which prohibited commercial banks from engaging directly in securities activities. The statutory definition of a "security" is an expansive one; see Huber (1989). According to the 1934 Securities Exchange Act, the term "security" means not only any stock, bond, debenture, and evidence of indebtedness, but also the "countless and variable schemes devised by those who seek the use of the money of others on the promise of profits." However, a general exception is made for situations where the *context* makes it inappropriate to treat an instrument as a security. For example, a loan participation purchased by a depository institution from another institution is not considered a security. The minimum consequence of concluding that an instrument is a security is that the antifraud provisions of the securities laws become applicable. In practice, therefore, the distinction between a loan and a security is driven largely by legal interpretation that cannot always be supported on economic grounds. With the dismantling of the Glass-Steagall Act, this distinction has become somewhat of a moot point.

Floating interest rates have been one of the most important innovations in bank lending since the advent of the term loan. Provisions for adjusting loan rates periodically give banks some protection against interest rate risk. By combining the advantages of term and short-term loans, floating rates have allowed banks to compete for a share of the business credit market—even in the face of increased competition from the commercial paper market and other nonbank credit suppliers. At the same time, floating rates have effected changes in the other terms and conditions of commercial lending. An unintended consequence has been the loss of some borrowers who switched from banks to the capital market to obtain longer-term debt with greater fixity in the borrowing rate.

Details of Loan Agreements: A loan agreement specifies the obligations of borrower and lender, makes certain warranties, and usually places certain controls and restrictions on the borrower. It states the amount to be borrowed, or *the principal.* The agreement also states *the maturity*: short-term (less than 1 year), intermediate-term (1 to 5 years), and long-term (greater than 5 years). *The pricing formula* also is stated. The interest rate may be a fixed or a floating rate. If the interest rate is floating, it may be "prime-plus" (for example, the prime rate plus 1 percent) or "times-prime" (for example, the prime rate times 1.05). Pricing might also be at a "transaction rate," that is, the bank agrees *ex ante* to a fixed mark-up over a current money rate (for example, T-bill, the negotiable CD rate, or the commercial paper rate). The agreement also states the closing fees to be paid when the loan gets funded. In a competitive situation this fee may be 0.25 percent to 0.375 percent, and higher in other situations. Also, a penalty or default rate of interest may be stipulated for late or early payments.

Although loan agreements usually are tailored to meet the requirements of specific situations, most contain certain standard provisions, which may be divided into three general categories: conditions precedent, warranties (also called representations), and covenants and events of default.

The *"conditions precedent"* section includes requirements the borrower must satisfy before the bank is legally obliged to fund the loan. These conditions may include specific business transactions that must be completed or events that must have occurred. Other standard items are the opinions of counsel, certificate of no defaults, the note, and resolutions of the borrower's board of directors authorizing the transaction.

The *"warranties"* section of the loan agreement contains information and assumptions about the borrower's legal status and creditworthiness. By executing the loan agreement, the borrower attests to the accuracy and truth of the information provided as of the date of execution. Misrepresentation constitutes an event of default. Principal warranties include the following:

- A warranty that all financial statements submitted to the lender are genuine and fairly represent the financial position of the borrower (that is, that no material adverse change has occurred).
- The borrower has a valid title to all assets.
- The borrower has complied with all federal, state, and municipal laws and is not involved in litigation.
- The borrower has filed all necessary tax returns and has paid all taxes due.
- No need for third-party consent.
- No violation of existing agreements.
- Collateral offered is owned by the borrower and is free of liens.

Covenants are a negotiated part of loan agreements. Warranties verify certain statements by the borrower at the date of execution of the loan agreement. Covenants carry forward the warranties and establish the borrower's ongoing obligation to maintain a certain status for the loan's duration. Covenants set minimum standards for a borrower's future conduct and performance and thereby accelerate the loan in the event of untoward developments. Violation of a covenant creates an *event of default* and gives the bank the right to "accelerate" the required repayment. We will have more to say about covenants in a later section of this chapter.

Informational Problems in Loan Contracts and the Importance of Loan Performance

Informational Problems

If there were no informational problems in loans, there might not be any profits for banks in lending. At one extreme, the costless availability of information obviates the need for banks and other financial intermediaries. At the other, costly customer-specific information provides an opportunity for banks to profitably process information and facilitate lending. In general, the less transparent is the credit information about a given borrower, the greater is the bank's ability to utilize its "uniqueness" and the higher is its profit potential. Thus, the paucity of good credit information in the public domain is a thing for banks to desire.[8] Since we have already discussed the informational problems addressed by banks (Chapters 2 and 3), we will merely review these here. The first problem is that the borrower is privately informed about its own credit risk. Unless the bank can elicit at least part of this information, market failure could result (recall the discussion of Akerlof in Chapter 1). We will see shortly that *credit analysis* helps the bank reduce its informational disadvantage vis-à-vis the borrower.

The other problem in lending is *moral hazard.* When the borrower takes a loan from the bank, it becomes an agent of the bank and is in a similar relationship with the bank as the shareholders of a firm are with bondholders. This agency problem is manifested in the borrower's desire to take on additional risk to the bank's detriment, as we saw in Chapter 1. Loan contracts are therefore designed to control the borrower's risk-taking propensity. To the extent that some preference for risk remains, the loan contract should also enable the bank to monitor the borrower and prevent actions that increase the risk of default. We will see how collateral, loan covenants, and other features of loan contracts can be structured to meet this important objective.

Figure 5.5 pictorially depicts the informational problems in loan contracts.

The Importance of Loan Performance

The bank's loan portfolio affects the financial health and viability of the bank. When bank stock prices decline, quite often most of the decline in bank stock prices is attributable to information releases about asset quality problems at banks.

8. Consider the following quote, *"Let us state a simple but often overlooked proposition: The health of a country's banking industry is inversely related to the speed and efficiency of information transfer,"* Sanford Rose, "Why Banks Make So Many Bad Loans," *American Banker*, June 19, 1990.

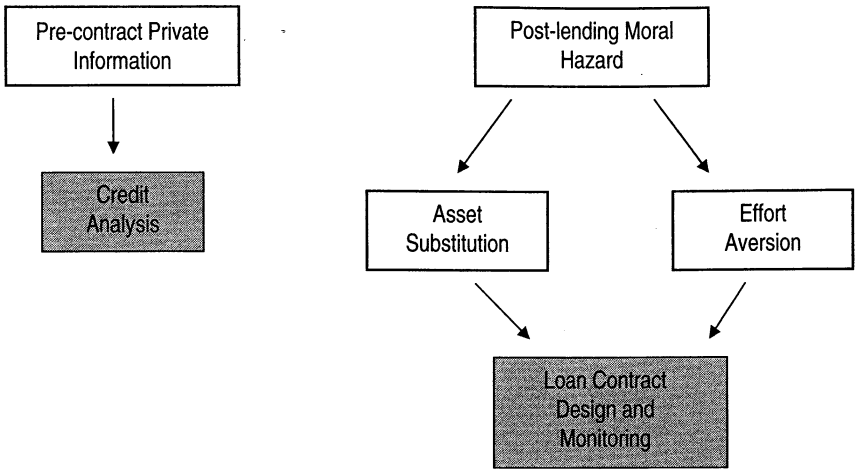


FIGURE 5.5 Information Problems in Loan Contracts

Loan losses can not only mean plunging stock prices, but can also spell trouble for top management at banks. It is well known that poorer corporate performance often precipitates a higher probability of CEO turnover.[9]

Loan Portfolio Diversification as a Risk Management Tool

The performance of a bank's loan portfolio often determines its financial performance. Although diversification is often a key to managing credit losses, many banks are constrained in their diversification efforts. For example, smaller banks often feel disadvantaged in their ability to diversify their credit risk by virtue of loan size limitations and geographic insularity. Moreover, these banks prefer limiting themselves to local markets because those are the markets they are most familiar with. This specialization-induced desire to stick to what is familiar leads to credit concentrations in banks, and this is typically reflected in the incidence of financial stress in periods of recession.

Despite regulatory attempts to encourage banks to diversify—by imposing limits on the maximum amount the bank can lend to a single borrower—the effects of lack of loan portfolio diversification can be clearly seen in the performance of banks in many regions. Banks in the United States have often displayed high performance correlation with banks that are similarly geographically situated. For example, when real estate values plunged in the 1980s in Texas, Oklahoma, and Louisiana, financial performance indicators for the banks in those states plunged as well.

It does appear, however, that the benefits of diversification have begun to more strongly influence banks' portfolio choices in the 1990s and post-2000, which coincides with the growing popularity of loan syndication, loan sales and securitization as diversification vehicles for banks. For example, although most United States community banks conduct much of their business in their own regions, there is recent evidence that these banks are able to withstand local economic downturns.[10] Moreover, in contrast to their relatively poor performance in the 1980s, small banks significantly

9. See, for example, Brickley (2003).

10. See John Hall and Timothy J. Yeager, *The Regional Economist*, "Does Relationship Banking Protect Small Banks From Economic Downturns?" *The Federal Reserve Bank of St Louis*, April 2002.

improved their performance in the 1990s. In fact, asset and deposit growth at small United States banks during the 1990s, when adjusted to account for the effects of mergers on measured growth, exceeded the growth at large banks.[11] In addition, small banks have also improved their profitability and survival rates. The FDIC reported that about 1,250 new community banks were established between 1992 and 2003, of which about 100 merged and about 1,100 remain independent, with only four having failed.

We now define terms that are routinely used in discussions of credit risk.

Interest Rate Spread: The difference between loan and deposit interest rates.

Provision for Loan Losses: A fraction of the loan principal earmarked by the bank as a buffer to absorb (expected) loan losses, and kept as part of the bank's capital.

Net Interest Spread After Provision: Interest rate spread after adjustment for taxes and subtraction of provision of loan losses.

Noninterest Income: Bank's income from activities other than lending, such as fees on cash management services, fees on contingent claims like loan commitments, letters of credit, and so on.

ROA: Bank's return on assets.

ROE: Bank's return on equity.

Nonperforming Loans/Reserves: Ratio of loans considered likely to default to the provision for loan losses.

Net Chargeoffs/Average Loans: Ratio of chargeoff of delinquent loans to the average loans extended by the bank.

Typically, interest rates are set such that interest rate spreads are higher for riskier loans. Banks also make higher provision for loan losses when the loans are riskier, and net chargeoffs/average loans also tend to be higher for such loans. Diversification can reduce the impact of losses in a particular loan class on the bank's overall net chargeoffs. Whether noninterest income, ROA and ROE are higher or lower for riskier loans depends on the degree of competition in that particular market and cannot be unambiguously stated *a priori.*

Despite the obvious gains from diversification, why are all banks not highly diversified? There are at least four reasons. First, there is the issue of limitations on the opportunity to diversify. Many banks feel "landlocked," constrained by geography to lend in limited markets. Second, lending opportunities typically arrive sequentially and unpredictably, so that forgoing a loan because of diversification concerns may be costly because a loan that offers better diversification potential may fail to materialize later. Third, banks are often constrained by regulations that mandate serving specific communities. For example, the *Community Reinvestment*

11. See William F. Bassett and Thomas F. Brady, "The Economic Performance of Small Banks, 1985–2000," *Federal Reserve Bulletin*, November 2001, pp. 719–728.

Act (CRA) requires a bank to lend to low-income borrowers in the community. This may interfere with diversification. Finally, cross-sectional reusability of information induces banks to specialize. For example, a bank that develops a special expertise in lending to auto parts manufacturers has a relative advantage in lending to this group, and it may wish to capitalize on this advantage by making such loans the focus of its loan portfolio. At best, therefore, banks tend to diversify within specialized areas of lending.

Credit Analysis: The Factors

Credit analysis examines factors that may lead to default in the repayment of a loan. The principal objective of credit analysis is to determine the ability and willingness of the borrower to repay the loan. The analysis looks at the borrower's past record (reputation) as well as its economic prospects. In most banks, this information is collected, analyzed, and stored by the credit department.

In analyzing a loan request, there are two important points to keep in mind. First, from an economic standpoint, assuming that the bank is the sole lender, it is the bank, not the borrower, that owns the asset financed with the loan. When the borrower takes a loan secured by the asset the loan is financing, it is merely purchasing a call option (as we saw in Chapter 1) from the bank. This option entitles the borrower to repurchase the asset from the bank should the value of the asset exceed the borrower's loan repayment obligation (the exercise price of the call option). The bank's loan granting decision and all of the actions it takes during the time the loan is outstanding should reflect this basic reality. Second, getting the borrower to repay the loan in today's legal environment is not always easy. Bankruptcy laws contain many provisions that protect borrowers, and these often make collection of debts potentially time-consuming and costly. Hence, one of the goals of credit analysis should be to uncover the likelihood of default as accurately as cost limitations will permit.

Traditional Factors Considered in Credit Analysis

Bank credit analysts have traditionally referred to the five Cs of credit analysis: capacity, character, capital, collateral, and conditions. Since "rules of thumb" are usually the distillate of accumulated experience, they should bear a relationship to theoretical prescriptions. We therefore, interpret each of these factors in terms of the underlying economics of bank lending. The discussion that follows is summarized in Figure 5.6.

(i) Capacity This refers to the borrower's legal and financial capacity to borrow. The first consideration in assessing a loan request is whether the person requesting the loan is legally capable of borrowing. For example, in the case of partnerships, it is important to know whether all the signing partners have the legal authority to borrow on behalf of the partnership. In the case of corporations, the bank should check the corporate charter and bylaws to determine who has the authority to borrow on the corporation's behalf.

Apart from legal considerations, capacity refers to the borrower's financial capability. Future cash flows are generally used to service the debt and therefore need to

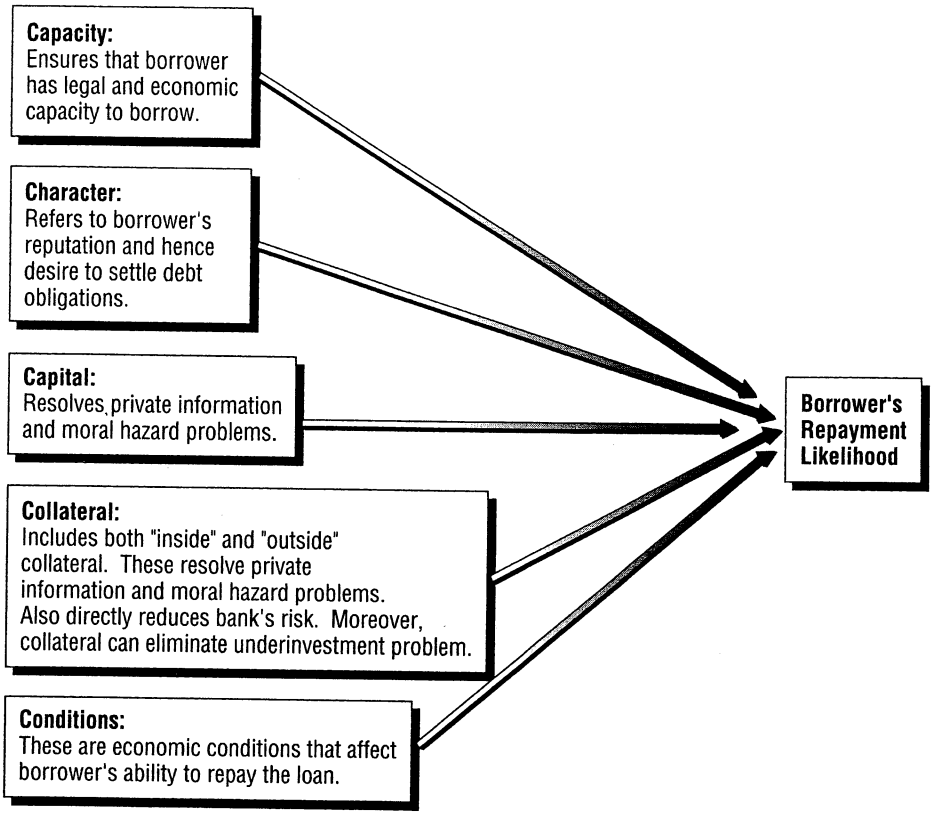


FIGURE 5.6 Pictorial Depiction of Factors Considered in Credit Analysis

be carefully estimated. Evaluating borrowers' future cash flows available to service the debt is a major part of any credit analysis. Sometimes, the bank may have to demand that the borrower subordinate the claims of others to ensure that the borrower has sufficient capacity to repay the bank. An example is a small firm that has borrowed significant amounts from its major shareholders.[12]

(ii) Character The concept of character embraces the borrower's ability to repay debts and the desire to settle all obligations within the terms of the contract. Judging character requires a careful examination of the borrower's past record in debt repayment and related behaviors. Including character in credit analysis makes sense because the better a borrower's credit reputation, the less incentive it has to default.[13] The reason is simple. Suppose a borrower knows that a single default will lead to denial of credit for a long time. The gain from defaulting is the amount the borrower does not repay the bank, but the gain from repayment is the net present value (NPV) of all the investment projects that might be financed with future bank loans; defaulting on this bank loan leads to a loss of that NPV. Clearly, this NPV increases as the interest rates on future loans decline. Further, the longer the borrower keeps repaying its loans, the better its credit reputation gets and the lower its future loan interest rates.[14] Hence, when the borrower acquires a good credit reputation, it perceives a

12. This is a case in which the bank may be successful in getting the borrower to subordinate the claims of earlier creditors. In general, this will be difficult as covenants on existing loans will generally prevent the borrower from taking such actions unilaterally.

13. This argument is formalized in Diamond (1989).

14. A better reputation leads to a lower interest rate because it becomes less likely that the borrower will default.

lower sequence of interest rates on its future loans than if it did not have that reputation. Consequently, the benefit of repaying the loan (or equivalently, the cost of defaulting) is greater for a borrower with a better reputation. To put it a little differently, the benefit of maintaining or building a reputation is greater the better the reputation is to start with. Hence, borrowers with better reputations (repayment records) tend to be better credit risks.

(iii) Capital How much equity capital (as a fraction of total assets) the borrower has invested in the firm is an important factor in the assessment of that firm's credit risk. There are two effects at work here. First, a higher amount of capital lessens the moral hazard problem. Second, the higher the capital, the better is the signal sent by the firm's owners about the confidence they have in the firm's future prospects. This helps to resolve the private information problem.

Example 5.1 Suppose you are a bank lending officer at the Midtown National Bank considering a loan request from Miller Manufacturing company for $1.05 million. The firm currently has $1 million in equity and its existing debt repayment obligation is $2 million. Assume that this equity is in the form of retained earnings invested in a noninterest-bearing account. The firm can invest the $1.05 million it will borrow from your bank in one of two projects (the bank cannot directly control which project the firm will invest in): A or B. Project A will yield a payoff of $2 million with probability 0.8 and $1 million with probability 0.2 at the end of the period. Project B will yield a payoff of $7 million with probability 0.2 and a payoff of zero with probability 0.8 at the end of the period. The firm's existing assets will yield a payoff of $3 million with probability 0.8 and a payoff of zero with probability 0.2 at the end of the period. The payoff on either project A or project B is statistically independent of the payoff on the firm's existing assets. These payoff distributions are common knowledge. For simplicity, there is no discounting and the bank loan you will make is subordinated to the firm's previous debt. Examine how Miller Manufacturing's behavior and the terms of lending change depending on whether or not it has the $1 million equity mentioned earlier.

Solution We solve this problem in four steps. First, we will assume that Miller Manufacturing has $1 million in equity. Then we will analyze the firm's expected profit from choosing project A, *assuming* that the bank prices the loan believing that project A will be chosen. Second, continuing to assume that Miller has $1 million in equity, we will analyze the borrower's expected profit from choosing project B assuming that the bank prices the loan believing that project A will be chosen. These two steps are needed to determine the appropriate Nash equilibrium in this problem, that is, a situation in which the bank prices the loan believing that Miller will choose project i (where i is either A or B) and Miller indeed chooses project i. With $1 million in equity, the Nash equilibrium involves the bank believing that project A will be chosen. Note the key role of the informational assumption that the bank cannot observe the borrower's choice of project. The third step is to assume that Miller has no equity capital and repeat Step 1. Finally, we repeat Step 2 with the assumption that Miller has no equity capital.

(*Continued*)

Step 1 Suppose first that the firm has the $1 million in equity mentioned earlier. Let's say that you, as the lending officer, assume that Miller will choose project A. Then, the *sum* of the cash flows from the project and Miller's existing assets has the following probability distribution.

TABLE 5.1 Probability Distribution of Total Cash Flows From Project A, Miller's Existing Assets and Equity

Total Cash Flow From Project A and Existing Assets Millions of $	Total Cash Flow With Retained Earnings Added in Millions of $	Probability
5	6	0.64
4	5	0.16
2	3	0.16
1	2	0.04

Since the repayment obligation on the senior debt is $2 million, the cash flow available to service the bank loan has the following probability distribution.

TABLE 5.2 Probability Distribution of Cash Flow Available to Service Bank Loan

Cash Flow Available Millions of $	Probability
4	0.64
3	0.16
1	0.16
0	0.04

You want to price this loan competitively because Miller has also been talking to your crosstown rival. At the same time, you do not want to lose money on this deal. From Table 5.2 you figure out that if the available cash flow is either $4 million or $3 million, Miller can fully repay the bank loan, whereas if the available cash flow is $1 million, then that is all your bank can collect. Thus, if you set the repayment obligation on your bank loan at $P million, your expected collection will be

$$(0.64 + 0.16)P + (0.16)1 + (0.04)0 = 0.8P + 0.16.$$

Since we've set the discount rate at zero, this expected payoff must equal the initial loan for your bank to just break even (the farthest you can go in competing for this borrower). That is,

$$1.05 = 0.8P + 0.16,$$

which means P = $1.1125 million, implying a loan interest rate of approximately 5.95 percent. The probability distribution of cash flows to Miller's shareholders is given in the table below.

TABLE 5.3 Probability Distribution of Net (Pretax) Cash Flow Accruing to Shareholders of Miller Manufacturing

Cash Flow Available Millions of $	Probability
2.8875	0.64
1.8875	0.16
0	0.16
0	0.04

Thus, the expected value of equity if Miller invests in project A is $0.64(2.8875) + 0.16(1.8875) = \$2.15$ million.

Step 2 Now suppose that Miller were to consider investing in project B after receiving a loan priced by you under the assumption that project A would be chosen. This is the standard moral hazard problem in bank lending that we discussed earlier, since project B is riskier for you as the lender. Then, proceeding in the same way that we did for project A, we see that the probability distribution of the cash flows accruing to the firm's shareholders is as follows: \$7.8875 million with probability 0.16, \$4.8875 million with probability 0.04, \$0.8875 million with probability 0.64, and zero with probability 0.16. Thus, the expected value of equity if the firm invests in project B is $0.16(7.8875) + 0.04(4.8875) + 0.64(0.8875) + 0.16(0) = \$2.0255$ million.

This means that Miller's shareholders prefer to invest in project A (assuming you price your loan as if project A will be selected) and you are safe in your assumption that project A will be chosen. It is, therefore, unnecessary to check what would happen if the bank were to assume that Miller will choose B. This is because there are two possibilities. Either Miller will choose A, so that it is not a Nash equilibrium for the bank to assume B will be chosen, or there is a Nash equilibrium in which Miller chooses B. But this Nash equilibrium is dominated by the one in which Miller chooses A in the sense that Miller is better off in the latter and the bank is indifferent. Thus, if your bank is to be competitive, you had better price the loan assuming that A will be chosen, since the loan price is lower in that case.

Step 3 Now we will see what would happen if Miller had no equity capital. In this case, if Miller selects project A, it has the following distribution for its total cash flow.

TABLE 5.4 Probability Distribution of Total Cash Flows from Project A and Miller's Existing Assets

Total Cash Flow Millions of $	Probability
5	0.64
4	0.16
2	0.16
1	0.04

Since the repayment obligation on senior debt is \$2 million, you calculate that to service the bank loan Miller will have \$3 million with probability 0.64, \$2 million with probability 0.16, and nothing with probability 0.2. Following the same logic as in the case with

(*Continued*)

$1 million in retained earnings, you now calculate that to permit the bank to just break even you must ask for a repayment obligation of $1.3125 million (you're assuming that Miller will invest in project A). With this, the net cash flow accruing to Miller's shareholders is $1.6875 million with probability 0.64, $0.6875 million with probability 0.16, and zero with probability 0.2. The expected value of equity is $1.19 million.

Step 4 After receiving such a loan, if Miller were to decide to opt for project B instead, we can follow the same steps as before to compute the expected value of equity as $1.2175 million. Thus, Miller will choose the riskier project B, and your assumption that it will select project A is incorrect. Indeed, if you were to (correctly) assume that project B will be chosen and price the loan accordingly, Miller's incentive to choose project B would be unaltered. This means that if Miller does not have sufficient equity capital, it may opt for riskier investments than it would if it had equity capital. Since you will anticipate this as a banker, you price the loan accordingly (that is, charge an appropriately higher interest rate on the loan). It is straightforward to verify that in this example Miller is better off retaining earnings in order to convince the lender that it will choose the safer project.

Capital helps to resolve moral hazard by imposing a greater loss on the borrower for poor project outcomes. This is because capital acts as the "first line of defense" against project losses and provides a cushion of protection for the lender. Without equity capital, the borrower knows that it has a valuable call option—if the project does poorly, the lender sustains the loss (the worst the borrower can do is to get nothing), whereas if the project does well, the lender gets only its contractual payment and the borrower earns a profit. With capital, the borrower's cost of pursuing risk is increased and the value of its call option is reduced. With sufficient equity capital, the lender can align the borrower's interest perfectly with its own. Interestingly, this means that the borrower is better off.[15]

The other function of capital is as an information communicator. The entrepreneur's own contribution of equity can signal the profitability of her project.[16] The standard argument relies on the entrepreneur being risk averse and is thus a little more complicated than an alternative line of reasoning that is developed in the example in the box below.[17]

Example 5.2 Suppose we have a firm that needs $150 to invest in a project that will yield a random payoff one period hence. The firm knows the probability distribution of the project's cash flow, but no one else does. All that others know is that the project can be type C or type D. If it is type C, then it will yield a cash flow of $300 with probability 0.8 and zero with probability of 0.2. If it is type D, the project will yield a

15. This is because of our assumption that the pricing of bank loans is competitive, so that the greater the equity capital possessed by the borrower, the better are its credit terms. Note that this provides an incentive for borrowers to accumulate equity capital.

16. See Leland and Pyle (1977).

17. This example is in the spirit of papers in the corporate finance literature that show a firm's choice of capital structure can signal its private information about its future prospects. See, for example, Ross (1977) and Shah and Thakor (1987).

cash flow of \$600 with probability 0.5 and zero with probability 0.5. For simplicity, suppose that interest and principal payments on debt are tax deductible and that the firm can raise equity capital (it currently has negligible equity capital on its books) from those who know the firm's cash flow distribution (for example, these may be managers who own stock). The firm currently has owners, but the book value of their equity is, for all practical purposes, zero. However, debt must be acquired in the form of a loan from a bank, which cannot tell whether the borrower has a type C or a type D project. The corporate tax rate applicable to the borrower is 30 percent. As a banker, how should you deal with such a borrower, assuming that the borrower is locked into either project C or project D and cannot choose its project?

Solution The key to resolving this informational asymmetry is to use capital as a signal. As a banker, the key is for you to recognize that the riskier borrower has a greater aversion to putting up equity capital because he has a greater likelihood of losing it. So, as a banker, you can offer the borrower two choices: (i) borrow the entire \$150 and repay P_D, or (ii) put up \$E in equity, borrow \$150 − E and repay P_c.

We solve this problem in three steps. First, we assume that the type-D borrower opts for choice (i), the type-C borrower opts for choice (ii), and the bank earns zero expected profit on each borrower. We then solve for P_D. We also solve for P_c, but it appears as a function of E. Step 2 involves solving for E. We do this by searching for the smallest value of E that ensures that the type-D borrower does not prefer its own contract (borrowing without putting up any equity) to that of the type-C borrower (putting up E in equity). Finally, the third step is to check that, with the value of E obtained from the previous step, the type-C borrower prefers his choice to that of the type-D borrower. Steps 2 and 3 therefore confirm the assumptions made in Step 1 about the project choices of borrowers.

Step 1 Now, if borrowers self-select so that only the type-D borrower takes (i) and only the type-C borrower takes (ii), then we can proceed as follows. Given that the bank must earn zero expected profit on each contract, and the repayment probability of the type-D borrower is 0.5, P_D must equal the expected value of the bank's repayment by the high-risk borrower, that is,

$$P_D \times 0.5 = 150$$
$$\text{or } P_D = \$300, \text{ an interest rate of 100 percent.}$$

Next, if only the low-risk borrower takes (ii), P_C must satisfy

$$0.8 \times P_C = 150 - E$$
$$\text{or } P_C = \frac{150 - E}{0.8}.$$

Step 2 We now solve for E. Note that E must ensure that the type-D borrower does not prefer the type-C borrower's contract to his own. Although there are many values of E for which this is true, there is only one value of E for which this is true and the value of the debt tax shield for the type-C borrower is maximized. This is the value of

(*Continued*)

E that is the smallest value such that the type-D borrower does not strictly prefer the type-C borrower's contract. That is, the NPV to the type-D borrower from misrepresenting [and choosing (ii)] is exactly equal to his NPV from telling the truth [and choosing (i)]. The type-D borrower's NPV from choosing (i) is

$$(600 - 300) \times 0.5 \times 0.7 = \$105$$

where 0.7 is one minus the tax rate. The type-D borrower's NPV from choosing (ii) is

$$\left(600 - \frac{150 - E}{0.8}\right) \times 0.5 \times 0.7 - E.$$

Equating the above NPV to \$105 yields E = \$70. Thus, the repayment obligation for the type-C borrower is $\frac{150 - 70}{0.8} = \$100$, or an interest rate of 25 percent.

Step 3 You can check that the type-C borrower will strictly prefer his contract to that of the type-D borrower. His NPV from (i) is

$$(300 - 300) \times 0.8 \times 0.7 = 0,$$

and his NPV from (ii) is

$$(300 - 100) \times 0.8 \times 0.7 - 70 = \$42.$$

Thus, the bank can offer two choices:

(i) Borrow the entire \$150 and repay \$300.
(ii) Put up \$70 in equity, borrow \$80, and repay \$100.

The key here is that the bank prices each loan based on the assumption that the borrower taking a particular loan has a particular project. If the borrower does in fact have that project, then the bank earns zero expected profit. The idea is for the bank to design the loan in such a way that *incentive compatibility* is assured. In other words, no borrower has an incentive to deviate from the loan contract "intended" for it by the bank. Incentive compatibility should obtain in a *Nash equilibrium*; the bank's assumptions about the association between the borrower's project and its loan contract choice must be correct in equilibrium.

In this example, capital serves as a signal of project quality. The borrower with the less risky type-C project signals its lower risk by funding two-thirds of the required investment with equity capital. For this, it is rewarded with a lower interest rate. Despite the obvious attractiveness of this lower interest rate, the high-risk borrower is unwilling to put up the equity necessary to be granted that rate. The intuition is as follows. Due to the tax deductibility of loan interest payments, the borrower desires as large a loan as possible, *regardless* of its project characteristics. The borrower also dislikes paying interest, regardless of its project characteristics. However, a higher interest rate is less onerous when the borrower has a risky project because the

likelihood of actually repaying the loan with interest is lower. To such a borrower then, the inducement of a lower interest rate in exchange for a higher capital requirement is less attractive than it is to a borrower with the safer project. It is the fact that the borrower's preferences over different capital requirement-interest rate combinations *depend* on its project characteristics that permit the bank to craft a self-selection mechanism that elicits the desired information.

It follows then that, all else remaining the same, the bank should charge an interest rate that is *inversely* related to the borrower's equity to total assets ratio. Less capitalized borrowers are more risky, not just because of the *direct effect* of capital in serving as a "first line of defense," but also because of its *indirect effect* in reducing the borrower's appetite for risk. In our examples, we imposed a zero-profit condition on the bank as a reflection of perfect competition in banking markets. This is an extreme representation of competition. In reality, banks earn profits, especially on borrowers about whom they possess credit information that is not publicly available. To the extent that banks charge higher interest rates to borrowers with lower equity positions, they may also be able to earn greater profit margins on these borrowers.[18] This can make the prospect of lending to highly leveraged (low equity) borrowers enticing for the bank, despite the higher risk involved. Indeed, such an incentive arises from the basic function of credit information production performed by banks (Chapter 3).

Banks can add highly leveraged loans to their portfolios by lending to companies that use the funds for leveraged buyouts, acquisitions, and recapitalizations. As our earlier discussions indicate, the yields on these highly leveraged transactions (HLTs) are higher than on other commercial loans. Since these higher yields compensate the bank for higher risks, higher *expected* profits for the bank are not necessarily implied. However, in many cases these borrowers also have few alternative sources of credit, so that banks can extract higher risk-adjusted profits from these borrowers. In addition, banks usually receive fees that vary from one to two percent of the principal amount committed.[19] HLT loans, however, are significantly more risky than average, and involve the moral hazards discussed earlier in this section.[20] This may be one reason why there has been a recent growth in the popularity of *reverse leveraged buyouts*, whereby firms reduce their debt/equity ratios by issuing equity to retire debt acquired during leveraged buyouts (LBOs). This would reduce moral hazard and benefit the firm.

(iv) Collateral Most commercial and consumer lending is secured with collateral.[21] Once a loan is secured by a specific asset that serves as collateral, the lender has first

18. This may also be because borrowers with lower equity capital levels may be less well known and have access to fewer credit sources, so that banks can earn higher quasi-monopoly rents by producing private information about them.

19. Usually, these loans are made under loan commitments, so that the fees are commitment fees.

20. An HLT loan may not only impose a higher expected loan loss for the bank but may also involve higher loan loss volatility (see Chapter 6).

21. For example, based on the Federal Reserve's Survey of Terms of Bank Lending, Boot, Thakor, and Udell (1991) report that, as of May 1988, 69.1 percent of bank loans were secured. See also Jimenez, Salas and Suarez (forthcoming).

claim to that asset in the event of default. There are two types of collateral: "inside" and "outside." Inside collateral consists of assets owned by the firm to which the loan is extended. Examples are accounts receivables, equipment, machinery, real estate, and inventory. Even if the bank extends an unsecured loan, it would have a claim, but not necessarily first claim, against these assets. As a general creditor, however, the value of the bank's claim would be ill-defined since, in the event of bankruptcy, the bank might be one among many unsecured creditors at the mercy of the bankruptcy court. On the other hand, if one of these assets is pledged as (inside) collateral, the bank would become the primary claimant to that asset.

Outside collateral consists of assets that the bank would never have a claim to unless they were specifically designated as collateral. A good example would be personal assets of the owner of the borrowing corporation or limited partnership.

Using collateral is not costless, however. Since the borrower may undertake actions that undermine the value of the collateral to the bank, ongoing monitoring of the collateral is required. Such monitoring costs are absorbed, at least in part, by the bank. Moreover, when collateral is transferred to the bank upon default, there are liquidation costs. These include the legal costs of ownership transfer as well as the bank's costs of initially carrying and then selling off the collateral.[22] From the borrower's standpoint, use of collateral makes subsequent borrowing more expensive since fewer assets are available to general creditors on that borrowing. Despite these costs, why is collateral so widely used?

There are at least three reasons for the popularity of secured lending. We discuss each now.

(a) Risk Reduction: An obvious reason to secure a loan is that it provides the lender greater protection against loss in the event of default. The bankruptcy code in the United States includes what is known as an "automatic stay," which freezes collection actions by creditors during bankruptcy proceedings. The idea is to provide the debtor with breathing room to put its house in order. The stay takes effect immediately upon the filing of a bankruptcy petition. However, the stay can be modified in favor of a creditor if there is "cause," including insufficient protection of the *secured* creditor's interest in that component of the debtor's property that serves as collateral. For example, suppose a bank has loaned $10 million to a firm that has just filed for reorganization under Chapter 11 of the bankruptcy code. Suppose that specific assets of the firm, currently worth $4 million, have been encumbered as inside collateral. Now, if these assets were to depreciate in value at the rate of $3,000 per month, for instance, the bankruptcy court might require the firm to set aside that amount each month to adequately protect the bank's claim. Thus, securing a loan reduces the creditor's risk in the event of bankruptcy.

(b) Signaling Instrument: Collateral can also convey valuable information to the bank. Although possible with inside collateral, the intuition comes through most clearly if one thinks of securing property as outside collateral. The logic is similar to that used in explaining the signaling role of equity capital. Within a class of borrowers that look equally risky to the bank even after all credit analysis is done, a borrower's willingness to offer collateral will be inversely related to its

22. By regulation, banks are required to liquidate such holdings within a certain time period after acquisition, unless the collateral is a permitted bank asset holding.

default risk on the loan.[23] The way the bank can induce a borrower to reveal its otherwise hidden risk is as follows. Suppose there are two indistinguishable borrowers, A and B. However, the bank suspects one may be riskier than the other, although it does not know which. The bank offers each borrower a choice of one contract from a pair consisting of a secured loan with an associated interest rate and an unsecured loan with a higher interest rate. Now suppose A is less risky than B. Then, A will prefer the secured loan for two reasons. First, its lower risk means that the likelihood of repaying interest is higher; hence, a lower interest rate is more appealing. Second, its lower risk means that the chance of defaulting and losing collateral to the bank is lower; hence, offering collateral is less onerous. By symmetric logic, we can see that B will prefer the unsecured loan. Getting A and B to sort themselves out like this requires, of course, that the two loan contracts offered are incentive compatible. The example in the box below shows how this can be done.

Example 5.3 Suppose that A's assets will be worth \$100 for sure at the end of the period. The value of B's end-of-period assets will be \$200 with probability 0.5 and zero with probability 0.5. The project (A or B) requires an investment of \$30 up front and the entire amount is borrowed from the bank. The bank is unable to distinguish between A and B. Assume that the single-period riskless interest rate is 10 percent and everybody is risk neutral. Assume that collateral worth \$1 to the borrower is worth only 90 cents to the bank. The difference of 10 cents on the dollar can be viewed as the bank's cost of taking possession of the collateral. These repossession costs have two sources. First, assets acquired from a delinquent borrower are often worth less piecemeal to the bank than they are to the borrowers as components of a productive whole. Thus, the mere act of liquidating collateral by removing it from the other assets of the firm is costly. Second, transferring control of assets from the borrower to the bank involves legal and other administrative costs. These costs are an important reason why so many bankers see the value of collateral largely in terms of its incentive effects. The problem is to determine how the bank can design a *pair* of loan contracts such that each borrower will be induced to truthfully reveal its privately known risk.

Solution Following the intuition discussed earlier, we will need to offer borrowers two contracts: a secured loan and an unsecured loan. These contracts should be designed so that A, the safe borrower, chooses the secured loan and B, the risky borrower, chooses the unsecured loan. We solve this problem in three steps. In the first step, we solve for the interest rate on the secured loan for the bank to break even. Second, we solve for the interest rate on the unsecured loan. In the third step, we solve for the amount of collateral on the secured loan that will deter the risky borrower from preferring the secured to the unsecured loan.

(*Continued*)

23. See Bester (1985), Besanko and Thakor (1987a, 1987b), and Chan and Thakor (1987) for theoretical models that demonstrate this. Empirical evidence on the signaling role of collateral is provided by Jimenez, Salas and Suarez (forthcoming).

Step 1 Since A will surely repay the loan, the interest rate on the secured loan, r_u, that allows the bank to just break even is the single-period riskless rate of 10 percent.

Step 2 On the other hand, the interest rate on the unsecured loan, r_u, should be set to satisfy the following zero profit condition for the bank

$$[0.5 \times (1 + r_u) \times 30]/[1.10] = 30 \qquad [5.1]$$

The left-hand side of (5.1) is the discounted present value of the bank's payoff. The promised repayment is $\$30(1 + r_u)$, but there is only a 0.5 probability that the bank will be repaid. Since the bank is risk neutral, it discounts at the riskless interest rate of 10 percent. For the bank to exactly break even, the discounted present value of its expected payoff should exactly equal the initial loan. Note that our approach is consistent with the notion that the bank owns the project and it has sold the borrower a call option on the collateral at a fixed exercise price of $30 \times (1 + r_u)$. When the project value exceeds this exercise price, the borrower exercises the option to repurchase the project; this happens in the successful state. If the project fails, the borrower lets its option expire unexercised and the bank retains a worthless project. Solving (5.1) gives $1 + r_u = 2.2$. Hence, the repayment obligation on the unsecured loan is $2.2 \times 30 = \$66$.

Step 3 Now we solve for the amount of collateral that will deter B from mimicking A and opting for the secured loan. The amount of collateral, C, that makes B indifferent between the secured and unsecured loans is the solution to the following equation

$$0.5 \times (200 - 66) = 0.5 \times (200 - 33) - 0.5 \times C. \qquad [5.2]$$

In (5.2), the left-hand side is the expected value of the borrower's cash flow, net of repaying the bank, if it takes the unsecured loan. The right-hand side is the expected value of its net cash inflow if it chooses the secured loan. Note that the interest rate on the secured loan is 10 percent (since the bank assumes this loan will be taken by the safe borrower), so that the repayment obligation is $1.10 \times 30 = \$33$. There is a 0.5 probability that the borrower will default and lose its collateral to the bank.

Solving (5.2) yields $C = \$33$. Thus, if the bank demands a collateral whose value to the borrower is at least as great as \$33, only A will choose the secured loan with an interest rate of 10 percent. (Note that A's net expected cash flow with the secured loan is $\$100 - \$33 = \$67$, whereas with the unsecured loan it is $\$100 - \$66 = \$34$). B will choose the unsecured loan with an interest rate of 120 percent. The bank can thus sort its borrowers according to risk. The outcome is a Nash equilibrium; the bank's beliefs about which borrower chooses which loan is confirmed by their behaviors.

You must have noticed that the bank's collateral repossession cost had no bearing on the outcome. The reason is that the secured loan to A is *riskless*, so that A would never surrender collateral to the bank. Since the Nash equilibrium separates perfectly—each borrower revealing its type in equilibrium—and involves B choosing the unsecured loan, the bank never actually takes possession of collateral in this

example. In reality, of course, few loans are riskless. With default risk in lending to A, then the bank's repossession cost would have entered the outcome since it would have affected the interest rate on the secured loan.

(c) **Moral Hazard:** Using collateral can help resolve a variety of moral hazard problems. The three we will discuss here are: asset substitution, underinvestment, and inadequate effort supply.

Asset Substitution: Because of the option nature of the bank loan, the borrower has an incentive to choose a riskier project after obtaining the loan. In a manner similar to capital, collateral can deter such risk-taking. For present purposes, think of security offered as outside collateral. Consider the following example.

Example 5.4 Suppose Brown Bakery needs a $100 loan to finance a project that will pay off next period. Brown can choose between two projects: S (safe) and R (risky). The bank knows this but is unable to directly control the borrower's choice of project. S will yield a payoff of $300 with probability 0.9 and nothing with probability 0.1, and R will yield a payoff of $400 with probability 0.6 and nothing with probability 0.4. Everybody is risk neutral and the riskless rate is 10 percent. How should the bank design its loan contract so that Brown will choose the safer project? Assume once again that collateral worth $1 to Brown is worth 90 cents to the bank.

Solution The idea is for the bank to make it in Brown's best interest to choose S. This is achieved by demanding that Brown put up sufficient collateral. Since collateral is surrendered to the bank upon default, it makes project failure costly to the borrower. Consequently, the borrower will wish to minimize the likelihood of failure by choosing S. The key assumption here is that the bank cannot *directly* control Brown's project choice. We proceed in four steps. First, we will assume that the bank offers Brown an unsecured loan, assuming that S will be chosen. We will show that this cannot be a Nash equilibrium because Brown will choose R. Second, we will let the bank assume that R will be chosen and compute the interest rate on the unsecured loan. It turns out this is a Nash equilibrium in that Brown chooses R when faced with such an unsecured loan. Third, we ask whether another Nash equilibrium is possible, say with a secured loan. We solve for the level of collateral that ensures that Brown does not (strictly) prefer R to S. We do this by equating Brown Bakery's expected profits from R and S, given a secured loan contract will indeed be acceptable to Brown Bakery and the bank. Finally, we verify that it is a Nash equilibrium for Brown to choose S.

Step 1 First suppose the bank offers Brown an unsecured loan at an interest rate r_u. If the bank assumes that Brown will choose S, then the interest rate, r_u^S, at which the bank just breaks even, is given by

$$[0.9 \times (1 + r_u^S \times 100)]/[1.10] = 100. \qquad [5.3]$$

(*Continued*)

Solving (5.3) yields $r_u^S = 22.22$ percent. Can this be a Nash equilibrium in the sense that Brown does indeed choose S? To answer this question, let us compute Brown's net expected payoffs under R and S. If Brown chooses S, its net expected payoff is

$$0.9[300 - (1.22 \times 100)] = \$160.20.$$

If it chooses R, its net expected payoff is

$$0.6(400 - 122) = \$166.8.$$

Hence, offering Brown an unsecured loan with an interest rate of 22 percent cannot be a Nash equilibrium since Brown will choose R instead of S, and the bank will make an expected loss on the loan since it assumed S would be chosen.

Step 2 Now suppose the bank assumes that R will be chosen. Then the interest rate, r_u^R, at which the bank just breaks even, is given by

$$0.6 \times (1 + r_u^R) \times 100]/[1.1] = 100. \qquad [5.4]$$

Solving (5.4) yields $r_u^R = 83.33$ percent. Now, confronted with this interest rate, if Brown chooses S, its net expected payoff is

$$0.9(300 - 186.33) = \$105.$$

If it chooses R, its net expected payoff is

$$0.6(400 - 183.33) = \$130.$$

So, Brown chooses R and this is a Nash equilibrium since the bank's belief is consistent with the borrower's behavior.

Step 3 But can we do better with *another* Nash equilibrium? Whenever we ask this question, it is natural to wonder who we are doing better for. Since the bank is assumed to earn zero expected profits in all scenarios, why should the bank care? The answer lies in competition. Recall that the zero expected profit condition is an analytical convenience. In practice we would expect the bank to earn at least a small profit. Remember too that this profit is in excess of the normal return on equity capital. Now, if the bank can design a contract that increases the borrower's expected profit without reducing the bank's, it can lure away this borrower from its competitors and build its "book" of business. Hence, competing banks should strive to give the borrower the best possible deal.

Suppose now that the bank offers Brown a secured loan instead. What you want to do as a banker is to figure out how much collateral to ask for in order to ensure that R will not be chosen. The level of collateral that leaves Brown indifferent between S and R satisfies the following equation.

$$\begin{aligned} &0.9[300 - (1 + r_s) \times 100] - 0.1 \times C \\ &\quad = 0.6[400 - (1 + r_s) \times 100] - 0.4C. \end{aligned} \qquad [5.5]$$

where r_s is the interest rate on the secured loan. We should first determine r_s. If the bank is successful in inducing Brown to choose S, then it should set r_s as follows to satisfy its zero profit condition

$$[0.9 \times (1 + r_s) \times 100 + 0.1 \times 0.9 \times C]/[1.1] = 100. \quad [5.6]$$

In (5.6), note that we have used the fact that a dollar of collateral is worth only 90 cents to the bank. Solving (5.6) yields

$$1 + r_s = (110 - 0.09C)/90. \quad [5.7]$$

Substituting (5.7) in (5.5) and solving for C yields C = $20,202. To avoid rounding off problems, suppose we take C = $20.21. Then substituting this in (5.7) gives us $1 + r_s = (110 - 1.8189)/90 = 1.2020$ or say $r_s = 20.21$ percent to make sure that rounding off does not leave the bank with negative expected profit.

Step 4 Now Brown's net expected payoff from choosing S is [from (5.5)] $159.79 and from choosing R it is [again from (5.5)] approximately $159.79. Hence, this is a Nash equilibrium in which Brown chooses S. Note that this equilibrium gives Brown a higher expected payoff than the previous Nash equilibrium ($130).[1] Thus, if this borrower comes to you and says that your crosstown rival has offered an unsecured loan at 83.33 percent interest, you could effectively counter by offering a secured loan that requires $20.21 of outside collateral and an interest rate of say 21 percent. With these terms, Brown Bakery will accept your loan and you will earn a profit.[2]

1. As noted in Chapter 1, there are often multiple Nash equilibria.

2. By this time, you may be wondering why a bank would ever make an *unsecured* loan. Note, however, that offering *both* secured and unsecured loans helps to resolve private information problems. Moreover, it is not *always* optimal to use outside collateral to resolve moral hazard. Indeed, in Example 5.4, if the payoff in the successful state for project R is $500 instead of $400, the best outcome is for the bank to offer an unsecured loan priced under the assumption that R will be chosen.

In this example, outside collateral was used since we assumed limited liability, that is, it would not be lost upon bankruptcy if it were not pledged. For somewhat different reasons, inside collateral can also deter asset substitution. By securing specific assets within the firm, creditors can ensure that these assets will not be replaced by those that increase the risk exposure of creditors. Since this reduction in asset substitution possibilities will be reflected in a better price for the firm's debt, the advantage of issuing secured debt accrues to the firm's shareholders.[24]

24. The argument that inside collateral can help in this way to resolve asset-substitution problems was made by Jackson and Kronman (1979) and Smith and Warner (1979).

Underinvestment: One manifestation of the divergence of interests between the borrower and the lender is in the borrower being unwilling to invest additional funds in a project even though doing so increases the total NPV of this project.[25] The intuition is simple. Suppose you own some real estate that was financed mainly with a bank loan; this real estate is currently worth \$1.5 million. You could spend an additional million dollars that would enhance the real estate's value by \$1.1 million. However, suppose that the present value of your repayment obligation to the bank is \$2 million. Then, although investing \$1 million yields an NPV of \$100,000 for the project as a whole, it is not a good idea for you, the owner/borrower. This is because you increase the present value of the cash flows accruing to you by \$(1.5 + 1.1)million − \$2 million = \$600,000, but it costs you \$1 million, that is, the investment has a *negative* NPV of \$400,000 to you (the borrower), but a positive NPV of \$100,000 to the borrower and lender considered jointly. The net effect is that the investment is passed up and firm value is sacrificed. This investment inefficiency arises from actions that are privately optimal for the borrowing firm's shareholders *ex post*. However, they pay a price for this *ex ante* since the lender anticipates such behavior and adjusts the terms of credit accordingly. How can we eliminate this form of moral hazard so that the *borrower* benefits *ex ante* through better credit terms?

One answer is to let the borrower *precommit* not to "underinvest" ex post. If the lender believes the borrower, the problem will have apparently been solved. However, such precommitment is *time inconsistent*. The lender knows that the borrower has every reason to break this promise when the opportunity presents itself. So it would be foolish for the lender to believe such a promise. Of course, loan covenants can be employed, with the lender monitoring compliance. However, as a practical matter, it is difficult to see how loan covenants could force a borrower to invest when it is disinclined to do so. This is because the lender typically does not "see" these investment opportunities unless the borrower decides to exploit them. Covenants are effective in *prohibiting* actions, but rarely succeed in forcing unobservable initiatives.

Secured debt can resolve this underinvestment problem.[26] The idea is as follows. Suppose that the firm needs additional financing to purchase an asset, and it can purchase this asset for less than its market value. Thus, the purchase is a positive NPV investment. Also suppose that the firm currently has risky unsecured debt outstanding and would not, without further incentive, purchase this asset because it would enhance the present value accruing to the firm's shareholders by less than the purchase price of the asset.

To solve this problem, suppose the firm issues new debt secured by the asset in question. Then, due to the "absolute priority" rule, the secured creditors have first claim to the asset in the event of bankruptcy, and the borrowing firm has essentially diverted (at least part of) the cash flows attributable to this asset to the new secured creditors and away from the old unsecured creditors. Since the new (secured) creditors pay a fair market value for the debt issued by the firm, the gains associated with diverting payoffs of the newly purchased asset away from the old (unsecured) creditors accrue to the borrowing firm's shareholders and increase their incentive to undertake the investment. The example in the box below illustrates how this works.

25. This underinvestment problem is discussed by Myers (1977).
26. This point was made by Stulz and Johnson (1985).

Example 5.5 Consider a firm, Johnson Supplies, that can invest \$100 at the start of the period (t = 0) in a project that will pay off at the end of the period (t = 1) \$400 if successful (state S_1) and zero if unsuccessful (state S_2). State S_1 occurs with probability 0.7. The initial \$100 financing comes from unsecured debt issued at t = 0. Before the end of the period, but *after* the initial financing is raised, the firm will have an opportunity to purchase an asset (call it A) for \$100. This asset will surely be worth \$120 at t = 1. Assume that Johnson cannot be forced to purchase this asset.[1] Compute Johnson's optimal financing strategy. Assume that everybody is risk neutral and that the riskless interest rate is 10 percent.

Solution We solve this problem in six steps. First, we assume that only unsecured debt can be offered and that the date-0 unsecured creditors will assume that Johnson will purchase A when available. We then compute the interest rate on the \$100 of (new) unsecured debt raised (*after* the initial financing) to purchase A. Second, we check if this can be a Nash equilibrium. We find that it is not, in that Johnson will *not* purchase A when burdened with the original unsecured debt. Third, we check if it is a Nash equilibrium for Johnson not to purchase A. That is, if the original creditors price their debt assuming that Johnson will not purchase A, will Johnson indeed not purchase A (since Johnson does not purchase A, we need not worry about the old creditors)? We find that this is a Nash equilibrium. Fourth, we introduce secured debt and compute the interest rates on the old unsecured and the new secured debt when all creditors assume that Johnson will purchase A when available. Fifth, we check if this is a Nash equilibrium. We find that it is a Nash equilibrium in that Johnson does purchase A and also wishes to issue secured debt to purchase A. Finally, in step 6 we conclude by indicating that the NPV to Johnson's shareholders is higher in the secured-debt Nash equilibrium than in the unsecured-debt Nash equilibrium when Johnson does not purchase A.

Step 1 First suppose that issuing secured debt is impossible. Thus, the \$100 financing required to purchase A in the future will have to be raised with either equity or unsecured debt. Since the basic argument follows in either case, let us assume that unsecured debt will be employed. As a start, suppose the unsecured creditors at t = 0 (call them C_{old}) assume that Johnson will purchase A when available. Use C_{new} to label the (new) unsecured creditors who provide the \$100 to buy A. Thus, at t = 1, the value of the firm will be \$520 (in state S_1) with probability 0.7 and \$120 (in state S_2) with probability 0.3. Assuming that all unsecured creditors have equal priority, C_{old} will be repaid in full in state S_1 and will receive \$60 in state S_2. The payoffs to C_{new} are identical. Hence, the loan interest rates on the credits provided by C_{old} and C_{new} will also be identical. Let r_a represent this interest rate. Then, if creditors provide fairly priced debt (that is, each creditor earns zero expected profit), r_a is obtained as a solution to the following equation

$$100 = [(1 + r_a) \times 100 \times 0.7 + 60 \times 0.3]/[1.1]. \qquad [5.8]$$

The left-hand side of (5.8) is the amount of debt financing. The right-hand side is the expected payoff to either C_{old} or C_{new}, discounted at the riskless rate of 10 percent. Solving (5.8) yields $r_a = 31.43$ percent. Thus, at t = 1 Johnson is obliged to repay \$131.43 to C_{old} and the same amount to C_{new}.

(*Continued*)

Step 2 The first question is: Can this be a Nash equilibrium? To answer this, we must find out whether C_{old}'s assumption that Johnson will purchase A is indeed correct. Now, if Johnson purchases A, the NPV accruing to its shareholders is

$$\frac{0.7 \times (520 - 262.86)}{1.1} = \$163.63.$$

Note that Johnson's shareholders receive a positive payoff only in state S_1, and this payoff is \$520(\$400 + \$120) minus two times \$131.43, where \$131.43 is what Johnson owes each group of unsecured creditors. If, on the other hand, Johnson does not purchase A, then the NPV accruing to its shareholders is

$$\frac{0.7 \times (400 - 131.43)}{1.1} = \$170.91.$$

Thus, Johnson will forgo the opportunity to purchase A even though its total NPV (\$120 − \$100/1.1 = \$18.18) to Johnson is positive. This means that it *cannot* be a Nash equilibrium for C_{old} to assume that Johnson will purchase A.

Step 3 So now suppose C_{old} assumes that Johnson will *not* purchase A. Then, the loan interest rate, r_b, is a solution to

$$[0.7 \times (1 + r_b) \times 100]/[1.1] = 100 \qquad [5.9]$$

Solving (5.9) yields $r_b = 57.143$ percent. It is simple to verify that, faced with this loan interest rate, Johnson will indeed choose *not* to purchase A. Thus, this is a Nash equilibrium, under the assumption that secured debt is impossible. The NPV accruing to Johnson's shareholders in this Nash equilibrium is given by

$$\frac{0.7 \times (400 - 157.143)}{1.1} = \$154.5.$$

Step 4 Imagine now that Johnson is free to finance A with secured debt. If Johnson chooses to do this, then the (secured) claim of C_{new} will be riskless since the minimum firm value (that prevails in state S_2) is \$120 (the value of A at t = 1), and C_{new} have first claim to this asset. Since the riskless rate is 10 percent, Johnson's repayment obligation on riskless debt will be \$110, and this can be covered from the value of this firm in state S_2. Now suppose C_{old} assumes that Johnson *will* purchase A when available. The loan interest rate, r_c, that C_{old} charges will then be a solution to

$$[0.7 \times (1 + r_c) \times 100 + 0.3 \times 10]/[1.1] = 100, \qquad [5.10]$$

where we recognize that C_{old} will be paid only \$10 in state S_2 since C_{old}'s claim is subordinated to that of C_{new}. Solving (5.10) gives us $r_c = 52.86$ percent. Johnson's total repayment obligation, therefore, is \$152.86 + \$110 = \$262.86.

Step 5 Is this a Nash equilibrium? Again, we consider Johnson's incentive to purchase A. If it purchases A, the NPV accruing to its shareholders is

$$\frac{0.7 \times (520 - 262.86)}{1.1} = \$163.63.$$

and if it does not purchase A, the NPV accruing to shareholders is

$$\frac{0.7 \times (400 - 152.86)}{1.1} = \$157.3.$$

Hence, Johnson will indeed purchase A (when C_{old} prices the loan assuming A will be purchased) and the conjecture of C_{old} about the firm's incentive to purchase A is supported by its behavior. To complete our verification that this is a Nash equilibrium, we must also make sure that Johnson will indeed wish to issue secured debt to purchase A. To check this, let us hold the fixed price of the loan given by C_{old}, so that the firm must repay \$152.86. If Johnson issues unsecured debt to purchase A, then C_{new} will ask for a loan interest rate of 31.43 percent [since they solve (5.8) to determine this loan interest rate], so that the NPV accruing to Johnson's shareholders is

$$\frac{0.7 \times [520 - (152.86 + 131.43)]}{1.1} = \$150.$$

Step 6 Thus, Johnson will indeed choose to finance A with secured debt. Moreover, the NPV to Johnson's shareholders in this Nash equilibrium (\$163.63) exceeds that in the previous Nash equilibrium when it could only finance the purchase of A with unsecured debt (\$154.5). Hence, it will *not* be in the interest of Johnson Supplies to precommit to never issue *secured* debt in the future through restrictive covenants written into its loan contract with C_{old}.

1. A simple way to ensure this is to assume that the opportunity to purchase the asset will arrive with some probability *less than one* and that creditors are unable to observe whether this opportunity has indeed arrived. This will not change the basic argument, but will complicate the numerical example a bit.

Apart from illustrating how secured debt can resolve the underinvestment problem, this example brings up an interesting point related to the design of covenants in loan contracts. It is sometimes believed that creditors wish to protect themselves against future expropriation by including loan covenants that prohibit the firm from issuing future debt that has a higher seniority claim against any subset of the firm's assets. When all is said and done, however, in a competitive market it is the *borrower* who decides what covenants to accept, since the lender can presumably adjust the price of the loan (to at least break even) depending on the covenants that the borrower is willing to accept. What our example shows is that it *may* be optimal for the borrower to leave itself the flexibility to avail of secured borrowing in the future in which the newly purchased assets are used as collateral, so that new creditors have the most senior claim to the assets.[27] This not only makes the borrower

27. Remember that in our example, C_{old} and C_{new} have equal seniority when the debt is unsecured, and C_{new} has higher seniority when it is secured. It should be noted, though, that our example does not show that it is optimal to issue new debt that has the senior-most claim against *all* of the firm's assets. Rather, the optimal new debt in the example is a prior claim against a subset of the assets and *no* claim against the remaining assets.

better off, but it even lowers the interest rate on the initial debt (C_{old} in our example). In our example, the interest rate on the loan provided by C_{old} is 57.143 percent when the issuance of debt of higher seniority in the future with respect to *any* asset is prohibited, and it is 52.86 percent when such issuance is permitted. The reason for this, of course, is that the ability to issue secured debt in the future resolves the underinvestment problem of debt.

Inadequate Effort Supply: Another moral hazard is that the borrower may expend insufficient effort in managing the firm when its assets are highly leveraged. Collateral can help to resolve this moral hazard problem, too. The following example uses outside collateral to illustrate the point.

Example 5.6 Consider an entrepreneur, Mr. David Barnes, who borrows \$100 at $t = 0$ (the start of the period) and invests the loan in a project that will pay off at $t = 1$ an amount \$300 in the successful state (state S_1) and nothing in the unsuccessful state (state S_2) for his start-up firm, Barnes Manufacturing. The probability of S_1 is p(e), where e is Mr. Barnes' effort in managing the project. Mr. Barnes can choose one of two effort levels: high (h) or low (ℓ). Mr. Barnes sustains a personal cost of \$40 to expend h and nothing if ℓ is chosen. Assume $p(h) = 0.8$ and $p(\ell) = 0.6$. Mr. Barnes has collateral available, but collateral worth \$1 to him is worth 90 cents to the bank. Assume that the bank cannot observe Mr. Barnes' choice of effort. The riskless interest rate is 10 percent. Compute the optimal loan contract.

Solution We want to show in this example that Mr. Barnes will work harder if the bank has loaned him \$100 with a secured debt contract. We will proceed in four steps. First, we will assume that the bank is restricted to offering an unsecured loan. We show that it is not a Nash equilibrium for Mr. Barnes to choose $e = h$. Second, continuing with the unsecured debt assumption, we show that it is a Nash equilibrium for Mr. Barnes to choose $e = \ell$, and for the bank to price its loan accordingly. Third, we introduce collateral and solve for the amount that makes Mr. Barnes indifferent between ℓ and h. We find that with this level of collateral it is indeed a Nash equilibrium for Mr. Barnes to choose h. Finally, in the fourth step, we check that Mr. Barnes himself is better off with secured debt, which serves as a precommitment that he will work harder.

Step 1 Suppose first that the bank restricts itself to offering an unsecured loan. If the bank assumes that Mr. Barnes will choose $e = h$, then the interest rate, r_h^u, it should charge on this unsecured loan to just break even satisfies

$$[0.8 \times (1 + r_h^u) \times 100/[1 + 0.10] = 100, \qquad [5.11]$$

which yields $r_h^u = 37.5$ percent. To check if this is a Nash equilibrium, we need to ask whether Mr. Barnes, faced with this loan contract, will indeed choose $e = h$. Mr. Barnes' expected payoff with $e = h$ is

$$0.8 \times (300 - 137.5) - 40 = 90,$$

whereas his expected payoff with $e = \ell$ is $0.6 \times (300 - 137.5) = 97.5$. Thus, this is not a Nash equilibrium since Mr. Barnes prefers $e = \ell$.

Step 2 It is, however, a Nash equilibrium for the bank to assume that Mr. Barnes will choose $e = \ell$, and price the unsecured loan accordingly. The loan interest rate, r_ℓ^u must satisfy

$$[0.6 \times (1 + r_\ell^u) \times 100]/[1.10] = 100, \qquad [5.12]$$

which yields $r_\ell^u = 83.33$ percent. Mr. Barnes' expected payoff with $e = h$ is $0.8 \times (300 - 183.33) - 40 = 53.34$. His expected payoff with $e = \ell$ is $0.6 \times (300 - 183.33) = 70.00$. Thus, it is a Nash equilibrium for the bank to price its unsecured loan assuming that Mr. Barnes will choose $e = \ell$.

Step 3 Now let us see if we can do better by using collateral. Let C be the collateral that leaves Mr. Barnes indifferent between choosing ℓ and h. Then r_ℓ^u and C must be related by the following equation

$$0.8 \times (1 + r_h^s) \times 100 + 0.2 \times 0.9C = 110. \qquad [5.13]$$

The left-hand side of (5.13) recognizes that the bank is repaid in full if the project is successful (this has probability 0.8) and only collects the collateral if the project fails (with probability 0.2). The value of the collateral to the bank is 0.9C. Solving (5.13) gives

$$1 + r_h^s = 1.375 - 0.00225C. \qquad [5.14]$$

Now, the amount of collateral needed to leave Mr. Barnes indifferent between ℓ and h is given by

$$\begin{aligned} &0.8 \times [300 - 100 \times (1.375 - 0.00225C)] - 0.2C - 40 \\ &= 0.6 \times [300 - 100 \times (1.375 - 0.00225C)] - 0.4C \end{aligned} \qquad [5.15]$$

Note that in (5.15) we have substituted for r_h^s using (5.14). Solving (5.15) yields $C = \$30.61$. Using this value of C in (5.14) gives $r_h^s = 30.613$ percent. To have Mr. Barnes strictly prefer h, suppose we choose $C = \$30.62$. Mr. Barnes' payoff if he chooses $e = h$ is now the left-hand side of (5.15) with $C = \$30.62$ and $r_h^s = 30.613$ percent. It is \$89,386. If Mr. Barnes chooses $e = \ell$, his expected payoff is the right-hand side of (5.15) and is given by \$89,384. Hence, Mr. Barnes prefers to choose h, and it is a Nash equilibrium for the bank to offer this secured loan on the assumption that Mr. Barnes will choose $e = h$.

Step 4 Note that Mr. Barnes' expected payoff in the Nash equilibrium with *unsecured* debt is \$70, whereas in the Nash equilibrium with secured debt it is \$89,384 (if Mr. Barnes chooses $e = \ell$) or \$89,386 (if Mr. Barnes chooses $e = h$). Thus, Mr. Barnes is better off by taking a secured loan, even though the use of collateral is dissipative.

We have discussed the various roles of collateral. The type and amount of collateral used will depend on which of these problems is dominant.[28] As mentioned earlier, using collateral can be costly, however, because of repossession costs. Additional costs are created because the quality of collateral must be appraised prior to making the loan and then monitored regularly during the life of the loan.[29] The reason for the appraisal and monitoring is that variations in the quality of a particular type of collateral across different borrowers may be quite large. For example, when collateral consists of accounts receivable, it will be of much higher quality if it is pledged by a borrower that has receivables due from well-capitalized companies with triple A ratings than if it is pledged by a borrower with receivables due from weak credit risks. Another example is *contract receivables*,[30] *whose* risk increases with volatility in business cycles. The point is that all collateral is not the same, and the deployment of collateral has various costs associated with it. These costs must be traded off against the potential benefits of collateral in deciding how to use collateral in lending. We turn now to the last of the "five Cs" of credit.

(v) Conditions By this we mean the economic conditions that affect the borrower's ability to repay the loan. Debts are repaid from four sources: income, sale of assets, sale of stock, and borrowing from another source. All of these should be assessed in determining the desirability, price, and other terms of the loan. The borrower's ability to generate income depends on: the selling prices of its goods, costs of inputs, competition, quality of goods and services, advertising effectiveness, and quality of management. Analysis of the borrower's financial statements as well as its management should inform the bank about the borrower's ability to create income.

In the Appendix, we discuss recent trends in credit analysis among banks. These highlight the increasingly sophisticated usage of computer technology in credit information processing.

Sources of Credit Information

The information used in underwriting credit is inherently costly and of uneven quality. The banker's critical skill in credit lies in assembling the most germane information at the lowest possible cost without violating legal requirements or social norms. This means identifying novel sources of information and using standard sources in clever ways. Following is a brief description of some of the standard sources of bank credit information, but we should emphasize that standard uses of

28. Empirical evidence on the relationship between collateral and borrower risk appears in Hester (1979), Orgler (1970), Morsman (1986), Berger and Udell (1990), Boot, Thakor, and Udell (1990), and Jimenez, Salas, and Saurina (forthcoming). These studies find that large prime borrowers are less likely to be asked to pledge collateral, whereas *observably* higher risk borrowers usually receive secured loans. (This is *not* inconsistent with our analysis that, among a group of *indistinguishable* borrowers, collateral can sort by inducing lower-risk borrowers to pledge more collateral). The finding that large, well-known borrowers are asked to pledge less collateral is also plausible since informational problems are likely to be less severe for such borrowers.

29. See, for example, Clarke (1987).

30. A "contract receivable" is an amount that a contractor is due to receive upon successful future completion of a contract. It involves chattel paper that shows the associated monetary obligations. Loans secured by contract receivables are often created when building or manufacturing contractors, dealers, or retailers need working capital.

standard sources is unlikely to produce anything better than average results. The clever use of credit information is a cultivated art form that distinguishes the successful lender from the pack.

Standard credit sources can be classified as: internal and external. By internal sources we mean those within the bank, and by external sources we mean all other.

Internal Sources

(i) **Interview with Applicant** The loan interview normally establishes the uses to which the borrowed funds will be put for the loan request and the conformity of the application with the bank's loan policies. For example, the bank's policy guidelines usually stipulate a minimum equity input by the/borrower, so that a violation of this guideline can be discussed with the borrower, leading perhaps to a smaller loan request. The loan interview is also used to judge intangibles related to the borrower's future repayment behavior. Moreover, it also provides the loan officer with an opportunity to advise the applicant about any additional financial information that might be needed for evaluating the application.

(ii) **Bank's Own Records** A bank normally maintains records of its depositors and borrowers. This source of information allows the bank to assess the borrower's past behavior.[31] For example, bank records will show the payment performance on previous loans, the balances carried in checking and savings accounts, and overdrawing patterns, if any. Even for applicants who have never been customers of the bank, the central file may contain some information if these applicants were solicited as potential customers.

External Sources

(i) **Borrower's Financial Statements** These are required of most borrowers. Audited statements are common requirements in commercial lending. Even in consumer lending, where loans are usually small, an applicant is normally asked to list what he/she owns, income and expenses, and outstanding debts.

(ii) **Credit Information Brokers** Information agencies or credit bureaus systematically collect financial information on potential borrowers and make it generally available at a price (recall Chapter 3). The most widely known is *Dun & Bradstreet* (D&B), which collects information on over 3 million businesses in the United States and Canada. D&B's *Business Information Report* provides information on the type of business, nature of ownership, composite credit rating, promptness with which the firm makes payment, sales, net worth, number of employees, general condition of the firm including information about its physical facilities, customer base, balance sheet information, the usual size of the firm's deposit balances, its payments record under loan agreements, and biographical information on principals. More detailed information can be found in D&B's *Key Account Report*. In *Dun's*

31. In Chapter 3, we pointed out that this may be an important advantage of banks in granting credit [see Fama(1980)].

Review, D&B also publishes information about financial ratios for a large number of industries.

Comparative financial information can also be found in the *Annual Statement Studies* published by *Robert Morris Associates*, a professional association of professional lenders. There are numerous other surveyors of credit information, specializing in consumer, business, and even governmental borrowers.

(iii) Other Banks Banks sometimes check with other banks that have had relationships with the loan applicant. They may also check with the firm's suppliers,[32] to learn how the firm pays its bills, and with the firm's customers to determine the quality of its products and the dependability of its service.

Analysis of Financial Statements

In evaluating the borrower's ability to service a loan, the bank will focus on the firm's internal sources for future generation of funds. These are: (i) net income, (ii) depreciation[33], (iii) reduction of accounts receivables, and (iv) reduction of inventories. To assess the potential of these cash flows, the bank examines the borrower's financial statements. However, financial statements are *noisy*. It is often necessary to work with audited statements that are months too old, along with unaudited interim statements that raise questions of authenticity. Even audited statements have their problems owing to the idiosyncracies of GAAP and the occasional lapses and professional compromises of auditors. These problems aside, financial statements value assets using nonmarket criteria such as book values, and income is distorted accordingly. Thus, financial statements should be interpreted with caution. An illustration is provided by the bursting of the stock market bubble in 2000 that was credited by some to a bond analyst raising questions about the credit worthiness of Amazon.com's debt based on accounting information not accurately reflecting cash flows for credit risk assessment purposes and concluding that Amazon's credit risk was higher than it seemed.[34]

Evaluation of the Balance Sheet

Assets

(a) Accounts Receivables: *Accounts receivables* are among the shortest maturity assets on the borrower's balance sheet and are typically seen as the major source of cash flows to service short-term loans. Standard analyses focus on the sizes, sources, and aging of accounts, as well as the extent to which the accounts receivables are actively managed and diversified. As with any other risky asset portfolio, diversification lowers risk. The bank may also wish to investigate the financial attributes of those who owe money to the borrower since these speak to the quality of the borrower's receivables. Credit bureaus are especially useful in

32. Another source of information about a potential borrower's suppliers is the *Credit Interchange Service* of the *National Association of Credit Management.*

33. Since depreciation is not a cash outflow but is subtracted in computing net income, it should be added back to arrive at cash flow.

34. Quite often, these issues are related to a divergence of accounting income from cash flows.

evaluating the quality of the borrower's receivables. Also, the current status or aging of receivables is a powerful indicator of their quality. For example, if a large fraction of receivables are 90 days or older and the convention is to pay in 30 days or less, the implications are transparent.

Not all borrowers need to be screened equally carefully. Relatively low-risk borrowers who may be close to qualifying for unsecured loans often fall under a "bulk" or "blanket-assignment" lending plan. For such borrowers, the bank may require only monthly borrowing-base certificates and aging or inventory listing, without maintaining active day-to-day control over collections. In the next risk category may be customers who keep good records and have a well-diversified accounts receivables portfolio. For such borrowers, the bank may impose additional reporting requirements, including detailed assignment, collection, and aging schedules. In the highest risk category are borrowers with weak balance sheets and inadequate working capital. Here the bank requires all standard reports plus copies of shipping documents, delivery receipts, and assigned invoices against which the bank will lend.[35] It is common for the bank to require such borrowers to remit collections directly to the bank in the form of checks "in kind." This is a way for the bank to exercise additional control. The bank might even mail invoices directly to the accounts in the borrower's accounts receivables portfolio, asking for payments to be made directly to the bank.[36]

(b) Contract Receivables: A borrower may be a contractor who has been engaged to perform some task in the future. Official recognition of this may appear in *chattel paper* that shows the monetary obligations of the party for whom this task is being performed. These monetary obligations are called contract receivables. Chattel paper often serves as collateral for a working capital loan. Contract receivables are riskier than accounts receivables since payment is conditional on the borrower's future performance. There is consequently a *double moral hazard*, one that the borrower may not successfully complete the contracted task and the other that the third party may not pay the borrower even if the task is successfully completed.[37] Thus, greater monitoring efforts are warranted for contracts receivables.

(c) Inventory: The age, liquidity, price stability, obsolescence, shrinkage, the adequacy of insurance coverage, the stage of processing, and the firm's method of inventory accounting are all issues in evaluating inventories.

As with any other form of collateral, the bank should be concerned about incentive effects as well as liquidation value. However, valuing partially processed inventories is difficult and a credit-analysis art form. Both raw materials and finished goods inventories are easier to value and have greater liquidity than partially processed goods. In many cases, raw material inventories have the broadest market and the lowest price volatility. As with other collateral, monitoring is crucial in that inventory stocks are constantly in flux, with potentially damaging consequences for the secured lender.

35. This procedure is called "ledgering" the accounts. See Clarke (1987).
36. This procedure is often referred to as handling borrowers on a "notification" basis.
37. With accounts receivables, you can see that only one of these two hazards is present.

(d) Fixed Assets: Normally, banks do not consider the sale of a fixed asset as a source of funds for loan repayment. However, surplus fixed assets can be occasional and strategic sources of cash flows. Whereas the main importance of fixed assets lies in their ability to *produce cash flows* and *not* in their resale value, business restructurings often generate surplus fixed assets whose expeditious sale can be value creating.

(e) Intangible Assets: These include trademarks, patents, copyrights, and goodwill. These assets are normally accorded little value by a bank because of their illiquidity and measurement errors. There are, of course, exceptions, but by and large bankers apply large discounts to such assets.

(f) Amounts Due: Banks often take a dim view of a firm's management if the firm's assets include amounts due from officers and employees. Amounts due create the suspicion of internal fraud and nepotism.

Liabilities and Net Worth

(a) Accounts Payable: The borrower's accounts payable should speak volumes to its bank. If the borrower does not pay its trade creditors timely, why should the bank expect to be treated differently? The bank should ascertain whether payables are in the form of notes since this may indicate that the firm has been denied trade credit. The bank should be similarly alarmed if the borrower has been asked by its suppliers for cash-on-delivery (COD) terms. In case the borrower owes money to its own shareholders or officers, the bank should demand explanation and may ask that such liabilities be subordinated to any bank loan. The bank should also review the amounts accrued for taxes and other expenses.

(b) Long-Term Liabilities: These consist of term loans, debentures, notes, mortgage loans, and other liabilities with maturities exceeding 1 year. The bank should be concerned with the *nature* and *maturity of* these obligations and the provisions that have been made for meeting the required payments. Their covenants may also be important for the bank considering a loan request. In particular, it is important to know whether the outstanding debt is secured and if so, which assets have already been pledged as collateral.

(c) Net Worth: The importance of equity capital to credit analysis is transparent, given our earlier discussions. However, *accounting* net worth is a particularly treacherous account because it is fraught with measurement errors. This item is the residual of assets and liabilities, with each asset and liability independently evaluated *with error*. Hence, the net worth compounds all of the errors embedded in the underlying accounts. If all assets and liabilities could be evaluated at market, the net worth should be the economic value of equity claims. However, with accounting distortions and other measurement errors, accounting net worth can be a hard-to-interpret residual.

(d) Contingent Liabilities: These are important because of their potential to become *actual* liabilities. If they do, they could seriously impair the debt-servicing capability of the borrower. Assessing the relevant probabilities and exposures may call for considerable information and sophistication. Moreover, such liabilities do not always appear in the body of the borrower's balance sheet. Even when

footnote disclosures reveal the borrower's exposure (maximum liability), the present value of the liability depends also on the unspecified contingencies and probabilities.

The Income Statement

Income statement analysis complements balance sheet analysis. Bankers tend to emphasize the balance sheet in evaluating short-term loans, but devote greater attention to the income statement for longer-maturity loans. Recall that the balance sheet measures *stocks*, whereas the income statement measures *flows*. Hence, by looking at past and present income statements, the bank should be able to learn something about the degree of stability in the borrower's cash flows. Of course, in determining cash flow trends, the bank should be careful to note possible changes in the borrower's accounting practices can obfuscate.

The bank will often use both the balance sheet and the income statement in its *ratio analysis*. Key financial ratios convey information about the firm's liquidity, stability, profitability, and cash flow prospects.

Basically, there are four types of ratios: liquidity, activity (or turnover), profitability, and financial leverage.

(i) Two measures of liquidity are commonly used:
current ratio = current assets/current liabilities,
quick ratio (or acid test ratio) $= \dfrac{\text{current assets} - \text{inventories}}{\text{current liabilities}}$.
By "current" we mean a duration of less than 1 year.

(ii) Activity ratios include the following:
Inventory turnover ratio = sales/inventory.
Average collection period (in days) = receivables/sales per day.
Total assets turnover = sales/total assets.
Fixed asset turnover = sales/net fixed assets.

(iii) There are also numerous profitability ratios. These include:
Profit margin on sales = net profit after taxes/sales.
Return on total assets = net profit after taxes/total assets.
Return on net worth = net profit after taxes/net worth.

(iv) The leverage ratio is defined as total debt/total assets.

Perhaps the two most important leverage ratios used by lenders are: pretax interest coverage and total debt to EBITDA[38]. Pretax interest coverage is defined as net income from continuing operations before taxes divided by reported gross interest expense. EBITDA is earnings before interest, taxes, depreciation and amortization. Figure 5.7 shows the behavior of these ratios through time for investment-grade U.S. corporate borrowers. It shows that the credit risk of these borrowers has been declining since 2002.

It is worth emphasizing that these ratios are usually expressed in terms of accounting values. Since bankers evaluate these ratios against peers, it is useful to

38. See Sufi (2006), who empirically shows the important of total debt/EBITDA.

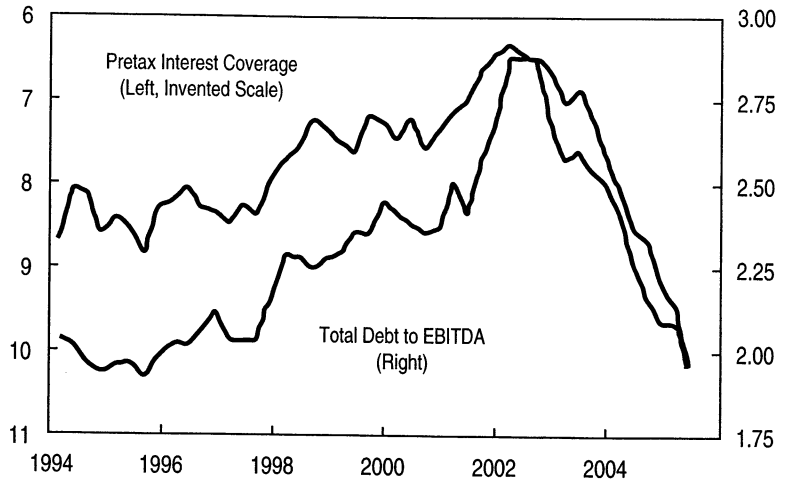


FIGURE 5.7 United States—Measures of Corporate Financial Performance for Investment Grade Corporate Borrowers (Ratio), 1994–2005
Note: Pretax interest coverage is net income from continuing operations before taxes divided by reported gross interest expense. Data are for industrial credits within the Citigroup BIG Credit Index.
Source: Citigroup.

remember that different firms may use different accounting methods. We provide a case at the end of this chapter that calls for ratio analysis as part of the credit evaluation process.

Loan Covenants

Covenants are special clauses designed to protect the bank and prohibit the borrower from taking actions that could adversely affect the likelihood of repayment. By agreeing to loan covenants that limit its actions, the borrower precommits to eschewing strategies that might expropriate wealth from the lender. The effect is to reduce the moral hazard faced by the lender and improve the terms of the loan agreement for the borrower. That is, loan covenants reduce the agency costs of debt and thereby benefit the borrower *ex ante*, and also the lender. Indeed, covenants make possible loans that would not otherwise be made at all. There is, of course, a limit to how restrictive a set of covenants the borrower will wish to accept. Restrictive covenants can make the loan reasonably safe for the lender but may deprive the borrower of valuable investment options and strategies.[39]

Loan covenants normally depend on the financial condition of the borrower, its investment opportunities, the track record of its management, and the lending

39. There may be circumstances in which restrictive loan covenants could perversely increase the likelihood of default by precluding actions the borrower could have taken to make both the bank and itself better off. For example, the purchase of new equipment by the borrower may be prohibited and yet the borrower's cash flows could be improved to such an extent by this purchase that the lender would be better off *ex post* if this covenant were relaxed. In such instances, the lender has an obvious incentive to renegotiate and relax the covenant [see Berlin and Mester (1992)]. However, if the lender is unsure of the borrower's motive for renegotiating and therefore uncertain of its potential benefit to the lender, it may refuse to renegotiate.

philosophy of the bank. Covenants are commonly classified into four kinds: affirmative covenants, restrictive clauses, negative covenants, and default provisions.

Affirmative Covenants

These are obligations imposed on the borrower. A commonly used covenant in this group is a requirement that the bank be periodically furnished with financial statements. The purpose, of course, is to permit the bank to keep track of the borrower's financial condition and enable preventive steps to be taken if trouble is indicated.

Another example is a requirement that the borrower maintain a minimum level of working capital. Banks will occasionally require the borrower to maintain a management acceptable to the bank. If management should change due to resignation, death, or other causes, the bank must approve the replacement.

Restrictive Clauses

These are designed to impose limits on the borrower's actions. A commonly used restrictive clause is one that limits the amount of dividends the borrower can pay its shareholders. The economic rationale for this covenant is transparent. A major concern for any creditor is the borrower's inclination to divert liquidity and net worth to shareholders rather than keep it within the firm to protect creditors.

It is also common for the bank to restrict salaries, bonuses, and advances to employees of the firm, as well as to limit specific types of investments such as purchases of fixed assets. The economic rationale for restrictions on investments is to protect creditors against asset substitutions that may reduce the value of the firm's debt. By purchasing a fixed asset, for example, the bank may be replacing cash on its balance sheet with an asset that will produce risky cash flows; this may increase the risk exposure of creditors.

Negative Covenants

While restrictive covenants limit certain actions, negative covenants prohibit them outright, absent the bank's consent. A common negative covenant is the *negative pledge clause*, usually found in unsecured loans. It prohibits the borrower from pledging any of its assets as security to other lenders. While the negative pledge clause is more common in unsecured loans, it is also encountered in secured loans. The banker may want to include this clause even though the bank's claim is protected with collateral because if the borrower defaults, the value of the collateral may be substantially diminished. In this case, bankruptcy law stipulates that for that portion of the bank's claim in excess of the value of the collateral, the bank has the same status as a general (unsecured) creditor. So the fewer the assets of the firm that are pledged for other loans, the greater is the share available to the bank in the event of bankruptcy.

There may also be prohibitions regarding mergers, consolidations, and sales of assets. The reason for this is that these developments can alter the firm's risk profile,

possibly to the creditor's detriment. It is also common for the bank to prohibit borrowers from making loans to others or guaranteeing the debts or other performances of others. Again, the economic rationale is clear. If the borrower were to do these things, it would assume additional credit risk on its account. By prohibiting such actions, the bank protects its own claim.

These are intended to make the entire loan immediately due and payable under certain conditions. Ordinarily, even though the bank has covenants that are intended to govern the borrower's behavior, violation need not automatically empower the bank to call the loan as long as scheduled payments are being made. However, some covenants will include an *acceleration clause* that specifies *events of default*. Effectively, violation of a covenant leading to any of the events of default automatically places the loan in default and full payment becomes due immediately. This permits the bank to take more timely actions than would be possible if it had to wait until a payment was missed. Acceleration clauses are often triggered by the following:

- Failure to make timely payments.
- Inaccuracy in representations and warranties.
- Violation of covenants.
- Bankruptcy, liquidation and/or appointment of a receiver.
- Entry of a judgment in excess of a specified amount.
- Impairment of collateral, invalidity of a guarantee and/or security agreement.
- Failure to pay other indebtedness when due or to perform under related agreements
  — Cross default.
  — Cross acceleration.
- Change of management or ownership.
- Expropriation of assets.

Any of the above may be considered an event of default, in which case the loan is accelerated and will lead to either renegotiation or default. In some cases, the loan agreement provides the borrower a period of time, referred to as a cure or grace period, to correct its default. If cured, the bank is then required to continue the loan. In the case where the default is not cured, the bank may terminate the lending relationship. The bank may also set off the borrower's deposits against its obligation to repay the loan and exercise its right to foreclose on collateral and even force the borrower into receivership. The cross default provision gives the bank the right to declare an event of default when the borrower is in default on another obligation. Though banks rarely exercise the right to accelerate loan repayment, having this right substantially strengthens a lender's position.

Other Parameters of the Loan Agreement

Loan agreements have many provisions other than amount and price that must be negotiated between the bank and the borrower. Some of the more important parameters of the loan agreement are:

- A *take-down schedule*: a time table for withdrawing funds from the bank.
- An *installment schedule*: a time table for repaying the interest, other charges, and principal.

- A *compensating balance requirement*: an obligation by the borrower to maintain deposits at the lending bank. (This requirement is usually stated in terms of the *average* deposit balance but may include minima as well.)
- A *prepayment provision*: a possible penalty for repaying a loan earlier than required.

The loan agreement also may contain provisions especially tailored to a specific situation. For example:

- The borrower agrees to sell, within the next 12 months, at public auction, or by any other commercially reasonable means, a commercial property owned by the borrower located at the corner of Oak and Spring Streets in Center City.
- The borrower agrees, within 180 days, to divest himself of his interest in a partnership known as Branson Truck Lines, and to apply any and all proceeds from the sale thereof to this loan.
- The borrower agrees to obtain, as soon as possible, and to assign to the bank, $100,000 of term life insurance.

It is worth keeping in mind that covenants, no matter how elaborate, can never anticipate all contingencies and prevent all disasters. For example, a borrower could have adequate liquidity as measured by its stock of working capital, and yet its actual liquidity position may be very poor because its accounts receivables portfolio is concentrated in a few high-credit-risk accounts. No loan covenants can replace vigilant and ongoing monitoring by the bank.

Conclusion

In this chapter we have examined the bank's spot lending decision. We have seen that a loan typically is an illiquid debt contract, without an active secondary market. The distinction between bank loans and traded bonds is significant on two grounds. First, trading tends to narrow informational gaps between borrowers and lenders, so that bank loans usually have less known about them than corporate bonds. Second, banks perform valuable screening services that overcome private information problems and postlending monitoring that resolves moral hazard problems. Thus, we should expect banks to lend to borrowers about whom less is known *a priori* and to those who have a rich set of investment opportunities so that moral hazard is a concern. This suggests a way to think about which borrowers approach banks and which go to the capital market (recall Chapter 3).

We have also discussed the design of loan contracts by banks in light of the informational problems they face. We have devoted considerable attention to the role of collateral and capital in overcoming these informational problems in traditional credit analysis.

Banks use a variety of internal and external information sources in order to perform the credit analysis needed to effectively screen borrowers. We have discussed these sources to highlight the potential impact of information availability on the bank's credit decision and its loan contract design. We hope that our discussions in this chapter have convinced you that the bank's lending decision is a complex one and expertise in credit analysis, loan contract design, and postlending monitoring is a

valuable resource. Hence, the uniqueness of a bank (recall Chapters 2 and 3). However, even the best experts cannot *always* effectively overcome informational problems in loan contracting. Sometimes these problems are insurmountable, and sometimes new information arrives that makes a previously negotiated loan contract inefficient. How banks deal with such situations is the subject of the next chapter.

Case Study Indiana Building Supplies, Inc.

The date is January 15, 2001. Alex Brown, vice president of the First National Bank of Bloomington (FNBB), was approached by Peter Willis, one of his loan officers who recently completed his training program at the bank after graduating with an MBA from a leading business school. Peter has been concerned about the financial ratios of one of FNBB's borrowers, Indiana Building Supplies, Inc. (IBS). The bank has installed a new software package to assist in its credit analysis, and this package monitors existing borrowers, alerting the bank to possible problems. This software package has indicated deterioration in some key financial ratios of IBS and has Peter worried about the likelihood that IBS will be able to repay the $473,000 it owes to FNBB by the due date of December 26, 2006.

Peter told Alex that he had run a special computer analysis on IBS about a month back and had noticed that some of the key financial ratios of the firm were trending downward. Peter based his assessment of IBS's ratios on the data provided in Tables 1 and 2. Not only were these ratios below the averages for the building supplies industry, but they were also at variance with the stipulations in the loan covenants negotiated between IBS and FNBB. Table 3 shows industry averages as well as loan covenant stipulations for key financial ratios for IBS. After his financial analysis, Peter contacted Bob Clemens, president of IBS, by phone and followed up with a letter providing details justifying his concerns. Clemens replied with a brief letter in which he conceded that some of the financial ratios had dipped below the levels specified in the loan covenants, but that there was no cause for alarm since the financial health of IBS was generally sound. Clemens pointed to the remarkable improvement in the firm's profit margin in 2005 relative to 2003 and 2004, and the fact that his return on net worth in 2005 was significantly above the industry average. When Peter called Clemens after receiving his reply, he explained to him that he was still concerned about the violations of ratio requirements in the covenants and wanted Clemens to send him data on the prices that IBS was charging customers for its finished goods. He also asked for (unaudited) quarterly financial statements on IBS.

Clemens seemed somewhat irritated by this request and reminded Peter that IBS had banked with FNBB for a long time and that Peter's predecessor had never been so picky with IBS even when it experienced substantially lower profit margins in 2003 and 2004. Nevertheless, he sent Peter the information he requested. When Peter analyzed this information, he found that IBS was charging higher prices than many of its competitors, especially those outside Indiana. Moreover, its quick ratio, current ratio, and its inventory turnover ratio all exhibited greater variations from quarter to quarter than the industry averages for these ratios.

IBS is a company that sells lumber products and a wide range of other building supplies in central and southern Indiana as well as in parts of Ohio and Missouri. Seasonal working capital needs as well as small capital equipment purchases have been financed primarily by loans from FNBB. IBS caters to basically two kinds of

customers: local customers in southern and central Indiana and those elsewhere. Demand from the Indiana customers is somewhat erratic, but because of their strong desire to purchase from local suppliers and IBS's long-standing reputation, their demand is less sensitive to price increases than the demand of the other customers. In the past, whenever costs of raw materials have escalated, Clemens has personally visited many of his local customers and explained to them that he needed to increase his prices to keep pace with rising costs. These efforts have been successful in convincing the Indiana customers not to switch to other suppliers. Clemens has been far less successful in passing along such price increases to other customers. They usually seem to be able to locate alternative sources of supply when IBS increases its prices.

Recently, David Klinghoffer, the chief financial officer (CFO) of IBS, has been urging Clemens to confine attention to IBS's "loyal" Indiana customers, and thereby reduce the marketing costs involved in reaching out-of-state customers. In the past, Clemens was reluctant to embrace this strategy because of the erratic nature of demand from Indiana customers. When IBS was price competitive, it could always count on a predictable level of demand from its Ohio and Missouri customers. Increased competition and higher costs, however, seriously damaged IBS's profit margins in 2003 and 2004 and persuaded Clemens to raise prices in 2005 to improve profitability. Klinghoffer, who had also been advocating higher prices, pointed out to Clemens with great delight that their strategy had been a smashing success and the firm had been more profitable in 2005 than it had ever been since 2000. Thus, both Klinghoffer and Clemens were dismayed by what they viewed as "senseless pestering" by Peter Willis.

The matter has now come before Alex Brown. Peter has pointed out to Alex that FNBB has an "acceleration clause" in its loan contract that empowers it to force IBS to repay its entire loan to FNBB immediately because of the violations of covenants. Alex was hesitant to do that and decided to call Clemens. When Alex advised him of the seriousness of the situation and the possibility that the bank would insist on immediate repayment of the entire loan unless some corrective action was taken, Clemens said it was likely that IBS would need an additional 1-year loan of about $200,000 (preferably at a 10 percent interest rate) to cover the amount payable on a note that was due to another creditor in a few weeks. He also requested FNBB to advise him regarding specific steps that the bank wanted IBS to take.

After hanging up the phone with Clemens, Alex asked Peter to bring him a detailed financial analysis of IBS, along with the specific reasons why Peter was so concerned. He also asked Peter to evaluate whether IBS's request for additional credit should be approved and to recommend specific steps IBS should be asked to take if the existing loan is not accelerated and new credit is granted. Alex wants Peter to pay particular attention to the fact that the "bottom line" *does* seem to indicate that IBS has done well in 2005, which makes Peter's worry somewhat anomalous.

Questions

Imagine that you are Peter Willis. Prepare a comprehensive ratio analysis for IBS. Should the bank call back the entire loan now? Why or why not? Should FNBB be worried or is Peter just overreacting? Is it possible for IBS to generate enough cash by year-end 2006 to make full repayment to FNBB? How valid are comparisons of IBS's financial ratios to the industry average?

TABLE 1 Indiana Building Supplies, Inc. Balance Sheet
Year Ended December 31

	2000	2003	2004	2005
Cash	$100,000	120,000	90,000	70,000
Accounts receivable	400,000	480,000	600,000	600,000
Inventory	500,000	550,000	800,000	900,000
Total Current Assets	$1,000,000	1,150,000	1,490,000	1,570,000
Land and building	100,000	90,000	217,000	221,000
Machinery	150,000	260,000	202,000	179,000
Other fixed assets	85,000	66,000	27,000	15,000
Total Assets	1,335,000	1,566,000	1,936,000	1,985,000
Notes payable, bank	47,000	53,000	110,000	473,000
Accounts and notes payable	156,000	171,500	233,800	319,000
Accruals	82,000	350,500	252,200	34,300
Total Current Liabilities	285,000	575,000	596,000	826,300
Mortgage	50,000	40,000	36,000	33,000
Common stock	900,000	900,000	1,150,000*	867,000**
Retained earnings	100,000	51,000	154,000	258,700
Total Liability and Equity	1,335,000	1,566,000	1,936,000	1,985,000

* The company issued common stock in 2004.
**In 2005 the company repurchased some stock, citing the unusually low market price of its stock.

TABLE 2 Indiana Building Supplies, Inc. Income Statement

	2000	2003	2004	2005
Net sales	$5,000,000	4,400,000	$5,600,000	$4,500,000
Cost of goods sold	4,000,200	3,400,000	4,500,000	3,500,000
Gross operating profit	$ 999,800	$1,000,000	$1,100,000	$1,000,000
General administration, selling, and interest expenses	521,467	582,000	849,667	519,000
Depreciation	80,000	105,000	80,000	72,000
Miscellaneous	65,000	93,000	77,000	71,500
Net income before taxes	333,333	220,000	93,333	337,500
Taxes (40%)	133,333	88,000	37,333	135,000
Net income	$ 200,000	$ 132,000	$ 56,000	$ 202,500

TABLE 3 Indiana Building Supplies, Inc.

	Ratios Specified in Loan Covenants	Industry Averages for 2005
Quick ratio	≥ 1.7	1.6
Current ratio	≥ 2.5	2.5
Inventory turnover ratio	≥ 9.00	8.5
Average collection period	NA	37 days
Fixed-asset turnover	NA	13.3
Total asset turnover	NA	3.00
Return on total assets	NA	9.5%
Return on net worth	NA	15%
Debt ratio	$\leq 38\%$	31%
Profit margin on sales	NA	3%

Notes: These figures are based on year-end figures taken from balance sheets and income statements of representative firms in the industry. These figures have been roughly constant for the past 5 years.

Review Questions

1. What are the different types of assets on a bank's balance sheet?
2. What is a "bank loan"? What are the different ways in which a bank can acquire loans?
3. Discuss the similarities and differences between loans and securities.
4. What are the major informational problems in loan contracts?
5. What is the purpose of credit analysis? Compare and contrast capital budgeting within a nonfinancial firm with credit analysis within a bank.
6. What are "the 5 Cs of credit"? What do we mean by a borrower's "character" and why is it important?
7. Can you explain intuitively why capital can resolve asset substitution moral hazard?
8. Discuss intuitively how capital can help the bank to resolve "adverse selection" problems. It would be useful to start out by explaining first what we mean by "adverse selection," and why it is a problem for the bank. Can you relate this role of capital in a bank loan contract to a venture capitalist's insistence on a minimum equity capital input by an entrepreneur seeking venture capital?
9. Please address the following questions:
   (a) What is a reverse leveraged buyout?
   (b) What are the main reasons why customers of banks become higher-quality credits after reverse LBOs?
   (c) Why are we observing such a large increase in reverse LBOs now?
10. What is the extent of secured lending among C&I loans? What are the two main types of collateral?
11. What are the costs of collateral? Why is "outside" collateral so popular despite these costs?

12. What is "underinvestment moral hazard"? Explain the intuition underlying the claim that collateral can attenuate this moral hazard. What are the implications of this for the design of bank loan covenants?
13. What is a "contract receivable"? Why is it usually more risky than an "accounts receivable"?
14. What are the main sources of credit information for banks in conducting credit analysis?
15. What is the role of ratio analysis in credit assessment? What are its limitations?
16. Overheard was the following conversation between two friends:
    Tom: I find it offensive that a bank would tell me what to do and what not to do when it makes me a loan. After all, I own the asset I'll buy with the loan because I have an *equity* stake in it. The bank is only lending me the money.
    Jack: That's nonsense, Tom! When you buy an asset with a bank loan, its the *bank* that owns the asset, and don't you forget it.
    What do *you* think? Explain your answer.
17. What are "affirmative covenants," "restrictive clauses," "negative covenants," and "default provisions"? Discuss the role of each in the design of credit contracts.
18. What are "expert systems" and what are banks attempting to achieve with them as part of credit analysis?
19. Consider a firm that has a bank loan outstanding that requires the firm to repay \$900 one period hence. The firm has \$300 in retained earnings that can either be paid out as a dividend to the firm's shareholders or invested in a project that will yield a single cash flow one period hence. The firm has a choice of investing in a safe project S, or a risky project R. The safe project will yield \$1,000 for sure one period hence, whereas the risky project will yield \$2,000 with probability 0.4 and nothing with probability 0.6. Assume that everybody is risk neutral and that the discount rate is zero. Which project has the higher *total* NPV for the firm? Which project will the firm choose, assuming that decisions are made to maximize shareholder wealth?
20. You are a bank loan officer. ABC Corporation has requested a \$2.1 million loan. The corporation has \$2 million in retained earnings and an existing debt obligation that calls for a repayment of \$4 million one period hence. The firm has existing assets that will be worth \$6 million with probability 0.7 and nothing with probability 0.3 one period hence. These are the future values of the assets in place if the firm does not make any investment at present. The firm also has the choice of investing in one of two mutually exclusive projects (A or B). Project A will yield \$4 million with probability 0.7 and \$2 million with probability 0.3 one period hence. Its cash flows are uncorrelated with (and in addition to) those from the assets in place. Project B will yield \$13 million with probability 0.2 and nothing with probability 0.8. Its cash flows are also uncorrelated with those from the assets in place. Assume that everybody is risk neutral and that there is no discounting. Moreover, ABC's existing debt has seniority over any new bank loan. Compute ABC's project choice and your pricing of the bank loan in two cases: (i) ABC has \$2 million in retained earnings that will be kept within the firm for one period, (ii) ABC has already announced that the retained earnings will be paid out as dividends right now and hence unavailable to augment ABC's cash flows one period

hence. Assume that your bank's cost of funds is zero and the bank is competitive (prices the loan to earn zero expected profit).

21. Consider a firm that needs \$350 to invest in a project that will yield a single cash flow one period hence. The firm knows the probability distribution of this cash flow, but no one else does. As a banker you only know that the firm is either low risk (L) or high risk (H). If it is L, then it will yield \$500 with probability 0.8 and nothing with probability 0.2 one period hence. If it is H, it will yield \$1,500 with probability 0.6 and nothing with probability 0.4 one period hence. The firm itself knows whether it is H or L. Assume that both the principal and interest repayments on any debt are tax deductible. The corporate tax rate applicable to this firm is 0.2. There is no equity capital on the firm's books at present, but it would raise equity if needed. The firm is locked into being either L or H, but as a banker you cannot tell which type it is. Assume everybody is risk neutral and that the discount rate (and the bank's cost of funds) is zero. Also, your bank is competitive (prices loans to earn zero expected profit). Construct a scheme consisting of two different loan contracts (one requiring the borrower to finance the project partly with equity capital and the other requiring no equity) such that the firm will truthfully reveal its private information by its choice of loan contract.
22. Consider a firm that can invest \$250 right now, at $t = 0$, in a project that will yield a single cash flow one period hence, at $t = 1$. This \$250 investment will be raised by issuing unsecured debt at $t = 0$. The project will yield \$500 with probability 0.8 and nothing with probability 0.2 at $t = 1$. Immediately after the initial investment but before the end of the period (say at $t = 1/2$), the firm can purchase another asset, call it A, for \$250 also. If purchased, A will yield a sure payoff of \$300 at $t = 1$. Those who lend the firm money at $t = 0$ cannot observe at $t = 1/2$ whether the firm had this investment opportunity. Everybody is risk neutral and the riskless rate is 12 percent. If you are the banker the firm has approached for a \$250 loan at $t = 0$, compute the price of your loan in two cases: (i) the firm can finance the acquisition of asset A with unsecured debt or not at all, and (ii) the firm can finance the acquisition of asset A with debt secured by the asset in question. Assume that in case (i), your bank (the initial lender) will have the same seniority as the new (unsecured) creditors who supply funds to purchase A. Your bank is competitive in loan pricing.
23. Given below is an excerpt from "A Friendly Conversation." Critique it.

Butterworth: I'll let that pass because I want to address your question, Mike. You know over 70 percent of business loans are secured, and collateral has some really beneficial incentive effects from the bank's standpoint. Moreover, it permits the bank to engage in creative loan-contract design that helps to resolve some thorny informational problems. It also leads to improved bank monitoring of borrowers, which is a key function associated with both secured and unsecured lending. To make a really long story short, I think that business lending is a key component of banks' activities. If regulation discourages this, then I think we'll have seriously weakened the financial intermediation process.

Moderator: If the role of banks in business lending were to diminish, what sort of losses to society do you foresee, Beth?

Butterworth: That's my favorite topic, Mike, so we could be here all night if I get going. But just briefly, I think that in the process of originating these loans, designing loan contracts, structuring covenants, including the crafting of collateral requirements, monitoring, and the restructuring of loans for borrowers in financial distress, banks have developed considerable expertise. It would be a shame if the financial system evolved in such a way that these skills would need to be relearned by others.

24. What is the "lending function" and how can it be decomposed? What is the usefulness of the decomposition?

Appendix 5.1 Trends in Credit Analysis

Banks are becoming increasingly sophisticated in credit analysis, relying more on computer-based statistical analysis of borrower attributes to determine the level of risk inherent in a particular loan. We will discuss two recent examples.

Illustration 1: Mellon Bank has installed computer software called the *Zeta Credit Scoring System* to analyze risk for private and commercial corporate clients.[1] This software has been developed by Zeta Services, Inc., Hoboken, N.J., which analyzes the financial condition of about 4,000 publicly owned firms and publishes quarterly reports for bankers. Mellon has begun using the Zeta Risk Control System both for assessing the credit risk of potential private loan customers and for monitoring existing borrowers. The system is also used by the Royal Bank of Canada.

The program produces a credit score that represents the probability that a company will stay in business and service its debt. Many banks, like Mellon, do not rely exclusively on one credit assessment. For example, Mellon has its own internally developed credit scoring system that evaluates loans. It then compares its own ratings to those yielded by the Zeta scoring system and devotes special attention to loans for which the two evaluations are strikingly different. Other banks may rely additionally on credit rating issued by the rating agencies like Moody's and Standard and Poor's. The objective, of course, is to improve the management of credit risk.

These credit scoring systems are essentially predictive models based on discriminant analysis. The purpose is to look at the data on numerous past borrowers and determine a relatively parsimonious set of variables that could have most accurately predicted which of these borrowers would default. For example, Altman (1968) provided the following formula

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \qquad [A.1]$$

where X_1 = working capital/total assets (in percentage),
X_2 = retained earnings/total assets (in percentage),
X_3 = earnings before interest and taxes/total assets (in percentage),
X_4 = market value of equity/book value of total debt (in percentage),
X_5 = sales/total assets (actual number).

1. See Gullo (1990).

Altman suggests that a Z value below 2.68 means that there is a high likelihood that the firm will go bankrupt. Since this early scoring model, numerous variants have appeared, but the idea is the same.

Illustration 2: Security Pacific Corporation has adopted a technology developed by the Department of Defense, called "neural networking." It is a branch of artificial intelligence that attempts to recreate the process by which the human brain learns.[2] The purpose of the program is to analyze risks in different types of loans. It is claimed that the neural network and the "expert systems" (a better-known branch of artificial intelligence) can solve problems that traditional number-crunching computer systems cannot.

Expert systems solve problems by utilizing the knowledge of experts in the form of "what if" statements. A neural network, on the other hand, solves problems without depending on the programmed knowledge of experts. The program is designed to "learn" and change the weights on different variables—that lead to a credit score—by detecting patterns. Neural networks are patterned on the neural connections in the brain. Their ability to learn and adapt makes neural networks appropriate for problems involving behavioral scoring and risk analysis. For example, suppose a neural network has been asked to analyze consumer mortgage loan applications. Then it will examine each variable and compare it with those in previous applications. Since the computer knows which of these previous applications were approved by the bank and which variables were weighted more heavily than others, it can compare a new application with its record of past applications and recommend a decision.

Expert systems first became popular in the mid-1980s, but as of this date only about half of the largest banks—which tend to be pioneers in the adoption of new technology—are using them. Neural networks are an even more recent adoption. Apart from Security Pacific, some other banks that are using this technology are Chase Manhattan Corporation, Manufacturers Hanover Trust Company, and Citigroup.

Limitations of Credit Scoring Models

While the use of computerized credit scoring models has grown significantly, these models are not without their shortcomings. A key shortcoming is that the estimates used in these models are based on data drawn solely from *extended* loans. Thus, these estimates suffer from *selection bias*.[3] Alternative approaches include those that rely on estimates derived from data that also include the characteristics of *rejected* applicants.[4]

References

Altman, Ed, "Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy," *Journal of Finance* 23, September 1968, 589–609.

2. See Layne (1990).
3. A thoughtful review of credit scoring models is provided by Hand (2001).
4. This is known as a process of "reject inference." See Kiefer and Larson (2004).

Ausubel, Lawrence, "The Failure of Competition in the Credit Card Market," Banking Research Center Working Paper No. 153, KGSM, Northwestern University, October 1990.

Bassett, William F., and Thomas F. Brady, "The Economic Performance of Small Banks, 1985–2000," *Federal Reserve Bulletin*, November 2001, 719–728.

Berger, Allen N., and Gregory F. Udell, "Collateral, Loan Quality, and Bank Risk," *Journal of Monetary Economics* 25, 1990, 21–42.

Berlin, Mitchell, and Loretta Mester, "Debt Covenants and Renegotiation," *Journal of Financial Intermediation* 2–3, June 1992, 95–133.

Besanko, David, and Anjan V. Thakor, "Collateral and Rationing: Sorting Equilibria in Monopolistic and Competitive Credit Markets," *International Economic Review* 28, October 1987b, 671–689.

———, "Competitive Equilibria in the Credit Market Under Asymmetric Information," *Journal of Economic Theory* 42, June 1987a, 167–182.

Bester, Helmut, "Screening vs. Rationing in Credit Markets with Imperfect Information," *American Economic Review* 75, 1985, 850–855.

Boot, Arnoud, Anjan V. Thakor, and Gregory F. Udell, "Secured Lending and Default Risk: Equilibrium Analysis and Monetary Policy Implications," *The Economic Journal* 101–406, May 1991, 458–472.

Boyd, John, and Mark Gertler, "U.S. Commercial Banking: Trends, Cycles, and Policy," *NBER Macroeconomics Annual*, 1993.

Brickley, James, "Empirical Research on CEO Turnover and Firm-Performance: A Discussion," *Journal of Accounting and Economics* 36, 2003, 227–233.

Chan, Yuk-Shee, and Anjan V. Thakor, "Collateral and Competitive Equilibria with Moral Hazard and Private Information," *Journal of Finance* 42, June 1987, 345–364.

Clarke, Peter S., "Collateral Lessons," *ABA Banking Journal*, November 1987, 68–70.

Diamond, Douglas, "Reputation Acquisition in Debt Markets," *Journal of Political Economy* 97, August 1989, 828–862.

Fama, Eugene, "What's Different about Banks?" *Journal of Monetary Economics* 15, 1985, 29–39.

Green, Richard, "Investment Incentives, Debt, and Warrants," *Journal of Financial Economics* 13, March 1984, 115–136.

Gullo, Karen, "Mellon Adds Credit-Score System," *American Banker*, March 7, 1990.

Hall, John, and Timothy J. Yeager, "Does 'Relationships Banking' Protect Small Banks from Economic Downturns?" *The Regional Economist*, Federal Reserve Bank of St. Louis, April 2002.

Hand, D.J., "Modelling Consumer Credit Risk," IMA *Journal of Management Mathematics* 12, 2001, 139–155.

Hester, Donald, "Customer Relationships and Terms of Loans: Evidence from a Pilot Survey," *Journal of Money, Credit and Banking* 11, 1979, 349–357.

Huber, Stephen K., *Bank Officer's Handbook of Government Regulation*, Second Edition, Warren, Gorham & Lamont, 1989.

Jackson, T.H., and A.T. Kronman, "Secured Financing and Priorities Among Creditors," *Yale Law Review* 88, 1979, 11–43.

Jimenez, Gabriel, Vicente Salas and Jesus Saurina, "The Determinants of Collateral," forthcoming, *Journal of Financial Economics.*

Kiefer, Nicholes, M., and C. Erick Larson, "Specification and Information Issues in Credit Scoring," *Economic and Policy Analysis Working Paper 2004*, Comptroller of the Currency, December 2004.

Kulkosky, Edward, "21% of Banking Assets Now Tied to Mortgages," *American Banker*, January 21, 1994.

Layne, Richard, "Security to Try Neural Networking," *American Banker*, July 3, 1990.

Leland, Hayne, and David Pyle, "Information Asymmetries, Financial Structure, and Financial Intermediaries," *Journal of Finance* 32, May 1977, 371–387.

Morsman, E. Jr., "Commercial Loan Structuring," *Journal of Commercial Bank Lending* 68–10, 1986, 2–20.

Myers, Stewart, "Determinants of Corporate Borrowing," *Journal of Financial Economics 5*, 1977, 147–175.

Orgler, Yair, "A Credit Scoring Model for Commercial Loans," *Journal of Money, Credit and Banking* 2, 1970, 435–445.

Rose, Sanford, "Why Banks Make So Many Bad Loans," *American Banker*, June 19, 1990.

Ross, Stephen, "The Determination of Financial Structure: The Incentive-Signalling Approach," *Bell Journal of Economics and Management Science*, Spring 1977, 23–40.

Shah, Salman, and Anjan V. Thakor, "Optimal Capital Structure and Project Financing," *Journal of Economic Theory* 42, August 1987, 209–243.

Sheshunoff, Alex, "Best of Times or Worst of Times?" *ABA Banking Journal*, July 1988, 25–37.

Smith, Clifford W., and Jerold B. Warner, "On Financial Contracting: An Analysis of Bond Covenants," *Journal of Financial Economics* 7, 1979, 117–161.

Stulz, Rene M., and Herb Johnson, "An Analysis of Secured Debt," *Journal of Financial Economics* 14–4, December 1985, 501–522.

Sufi, Amir, "The Real Effects of Debt Certification: Evidence from the Introduction of Bank Loan Ratings," Working Paper University of Chicago, April 2006.

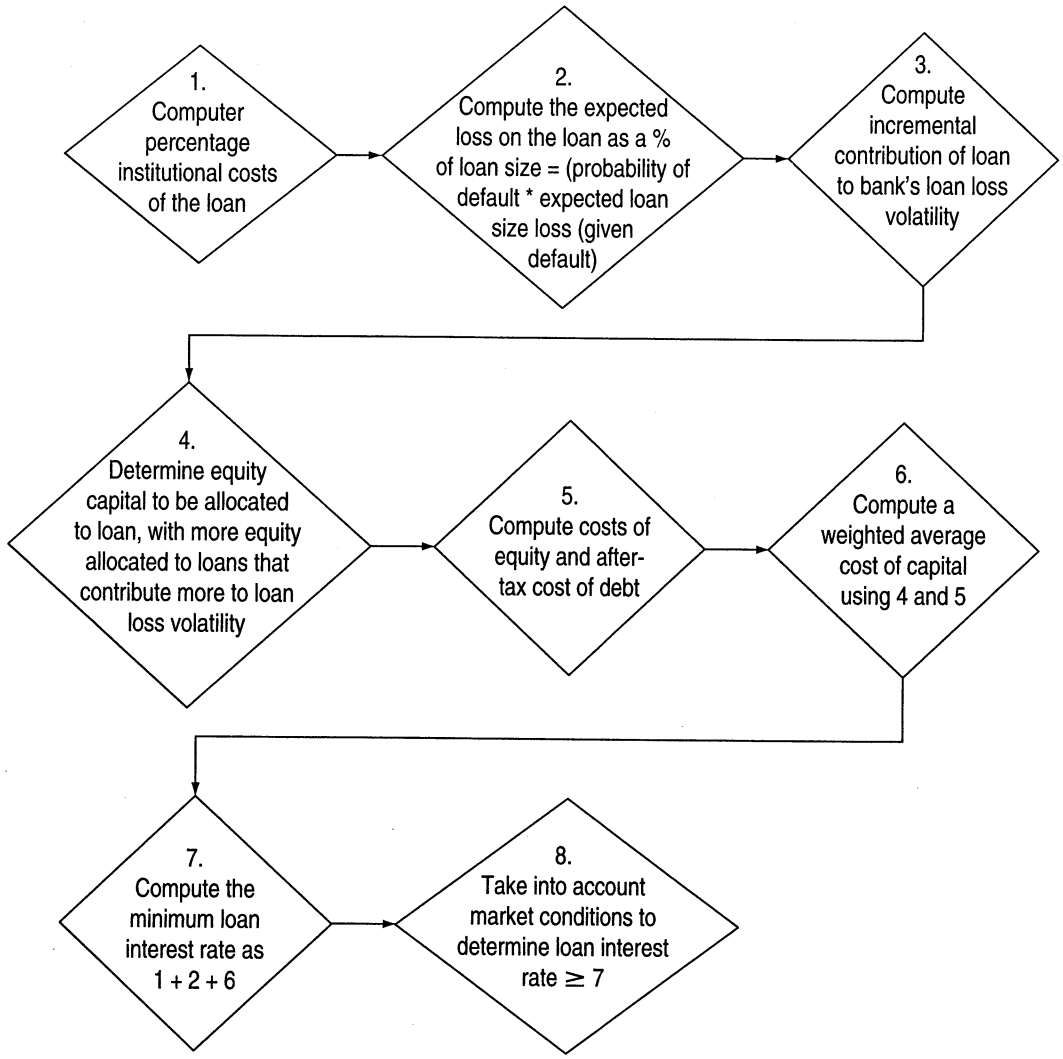


FIGURE 6.4 Loan Interest Rate Determination

Credit Rationing

Credit rationing is defined as a situation in which a lender refuses to extend credit to a borrower at the price *posted by the lender* for that borrower class. Credit rationing is *not* a phenomenon whereby a potential borrower refuses to accept credit because the price is "unfair" or too high. The essential point is that credit is denied at a price selected by the lender itself. Even if the borrower offers a higher interest rate than that asked for by the lender, a loan is refused by the lender.

Credit rationing is a puzzling practice.[5] When credit is rationed, there is an unsatisfied demand for credit at the price posted by the bank, that is, credit demand exceeds supply at that price. Conventional economic theory, or just plain common sense, suggests that the bank could increase its profits by increasing the price of credit. If the supply function for credit is upward sloping and the demand function is downward sloping, as shown in Figure 6.5, then this should bring about the usual equilibrium in which demand and supply are equated. Since the bank is supplying

5. Included in credit rationing is the practice of "redlining," which involves the lender refusing to extend the credit based on considerations of race, gender, and so on. This is illegal and is not the focus of our discussion.

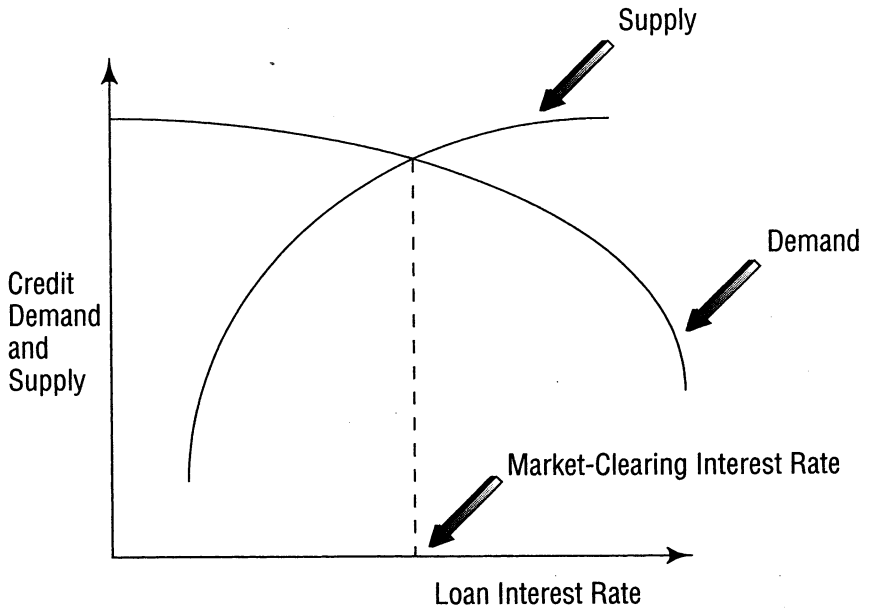


FIGURE 6.5 The Demand and Supply for Credit

more credit and at a higher price, its profit should be greater. Thus it seems irrational for profit-maximizing banks to ration credit.[6] Is it?

While it is conceivable that banks forgo profitable lending opportunities, it seems implausible. We thus ask whether it is rational for a profit-maximizing bank to ration credit.

Why Should We Be Interested in Credit Rationing?

It is believed that a fall in the money supply restricts spending. This could happen even if the fall in the money supply caused only a small increase in interest rates, or if spending is not curtailed by an interest rate increase. The reason is that a fall in the money supply would leave banks with less to lend, forcing them to reduce their lending, even if customers did not reduce their loan demand. Thus, spending was viewed as being constrained by the availability of credit to banks, and this credit was allocated to customers through nonprice means such as credit rationing. This argument, popularly known as the "availability doctrine," suggested an alternative transmission channel for monetary policy that was based in an important way on the monetary policy argument.

There are two reasons why we should be interested in studying credit rationing in connection with monetary policy. First, with credit rationing, monetary policy can be effective influencing aggregate investment by corporations even with little variation in interest rates. That is, if the Federal Reserve feels that inflationary pressures need to be abated by curtailing spending, it could cause a slowdown of the economy without major changes in interest rates. This could be achieved by reducing the liquidity of banks, which in turn could lead to reduced bank lending due to credit rationing, even if investment demand by corporations was unchanged. Thus, the effectiveness of monetary policy would have *not* been empirically documented. An important implication of this is that in the presence of credit rationing, the monetary policy options of inducing increased interest rates through a higher discount window borrowing rate

6. Samuelson (1952) was the first to suggest this.

and of reducing the amount of credit available through open market operations (bond sales) are not necessarily equivalent. Credit can be reduced even if investment demand is insensitive to monetary policy manipulations.

Second, it has been empirically found that a more stringent monetary policy does not affect all borrowers equally. Thus, if credit rationing is better understood with respect to the identities of those who are rationed, we may be able to better predict the effects of a restrictive monetary policy.[7]

Why Is There Credit Rationing?

In order to understand why a profit-maximizing bank might ration credit, we need to examine the conditions under which it would not be optimal for the bank to increase its loan interest rate when faced with excess demand for credit. It is difficult to see why banks would do this if they had as much information as the borrower. If the bank was perfectly informed, it could always set an appropriate risk-adjusted price and lend accordingly.

However, in a world of asymmetric information, credit rationing can be an optimal strategy for a profit-maximizing bank. The explanation turns upon two types of information hurdles.[8] First, a bank may not be able to distinguish perfectly between borrowers with different credit risks, even after it has analyzed each borrower's financial information. This is called the *precontract private information* problem. Even if the bank knows the *average* riskiness of borrowers within a given risk classification, it may not be able to identify individual risks [recall the Akerlof (1970) discussion in Chapter 1]. The bank will, therefore, charge a common price to all within the risk class, so that some borrowers are subsidizing others. A second problem is that the bank may not be able to completely control the borrower's actions. The borrower may thus be able to increase project risk, either through its choice of projects or through its expenditure of effort, without detection by the bank.

Now imagine that a loan interest rate is announced by the bank for a particular risk class, and at that interest rate there is an excess demand for loans by borrowers in that risk class. What would happen if the bank chose to increase the loan interest rate? One possibility is *adverse selection*. Safer borrowers within the given risk classification may be unwilling to borrow at the higher interest rate, so that the mix of borrowers within the pool becomes riskier. If this happens, the bank's expected profit could actually be *lower* at the higher interest rate; we provide a simple numerical example below to illustrate. A second possibility is that an increase in the loan interest rate could *worsen* the moral hazard problem. That is, those borrowers within the pool who have some latitude in their investment decisions may choose riskier projects at the higher loan interest rate. This again could mean a lower expected profit for the bank at the higher loan interest rate. Thus, the bank may conclude that increasing the loan interest rate is not worthwhile since its expected profit is maximized at an interest rate at which credit demand exceeds supply.[9] Figure 6.6 depicts this graphically.

7. Some evidence suggesting rationing is provided in Jaffee and Modigliani (1969).

8. What follows is an adaptation of Stiglitz and Weiss (1981).

9. That is, suppose r is the loan interest rate, C is the bank's per dollar cost of funds and θ is the repayment probability. Then the bank's expected return per dollar loaned is $\rho = [1 + r]\theta - C$. The point is that θ cannot be taken as being unaffected by r. As r is raised, θ falls. Assuming that θ is a decreasing and concave function of r (that is, $\partial\theta/\partial r < 0$, $\partial^2\theta/\partial r^2 < 0$), we see that the function $\rho(r) = [1 + r]\theta(r) - C$ attains a unique maximum with respect to r.

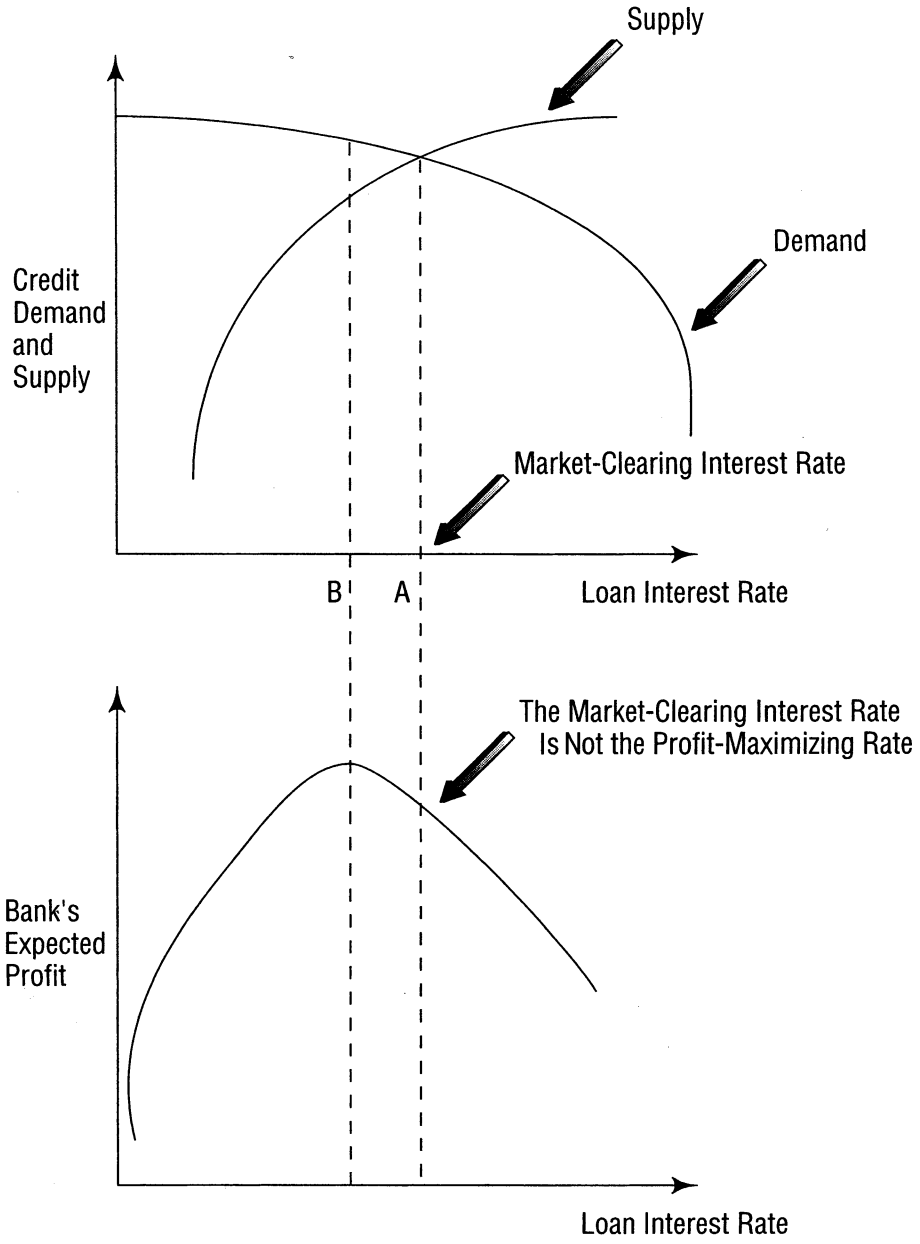


FIGURE 6.6 Credit Rationing

We now provide numerical examples to illustrate these concepts. We will first focus on the *adverse selection* problem, ignoring moral hazard for the moment.

Example 6.2 Suppose that you are the loan officer for the Midtown Community Bank and you know that within a particular risk class, there are two types of borrowers: low-risk borrowers and high-risk borrowers. However, you cannot distinguish between them.

You believe that the probability is 0.5 that a randomly chosen borrower is low risk and 0.5 that the borrower is high risk. There are 1,000 potential loan applications of each type within this risk class. Each applicant would like a loan of $100. The low-risk borrower will invest this loan in a project that one period hence will yield $130 with probability 0.9 and nothing with probability 0.1. The high-risk borrower will invest the loan in a project that will yield $135 with probability 0.8 and nothing with

(*Continued*)

probability 0.2 one period hence. Midtown Community Bank is a monopolist with respect to these borrowers.[1] Assuming that the only pricing instrument available is the loan interest rate, how should you price a loan to a borrower in this risk class so as to maximize the bank's expected profit? You have only $100,000 available to lend and the junior lending officer who reports to you has advised you that 2,000 loan applications were received when it was announced that the bank would charge an interest rate of 29 percent. The current riskless rate is 5 percent. Assume that a borrower must have at least 1 dollar of net profit in the successful state in order to apply for a bank loan,[2] and that there is universal risk neutrality.

Solution This example shows how informational considerations can impart rigidity to the bank's loan interest rate. To show this, we proceed in three steps. First, we will compute Midtown Community Bank's expected profit if it charges a rate of interest of 29 percent and is forced to randomly ration half its loan applicants (because all potential borrowers apply). Second, we calculate Midtown's expected profit if it charges a rate higher than 29 percent. In this case, the low-risk borrowers drop out, so that the bank lends only to the high-risk borrowers. Finally, in the third step, we compare the bank's expected profits from the first two steps and show that Midtown Community Bank's expected profit is maximized by setting the loan interest rate at 29 percent and randomly rationing half its credit applicants. The key to this finding is that the bank cannot distinguish between the low- and high-risk borrower.

Step 1 Clearly, if you charge an interest rate of 29 percent, you will have to ration credit since you can lend only $100,000 to this group of borrowers and the demand is for $200,000. Now, the *maximum* interest rate that your bank can charge without losing the low-risk borrowers is 29 percent. At this interest rate, the net profit of the low-risk borrower in the successful state is

$$130 - 129 = \$1,$$

because the repayment obligation is $129. Clearly, the high-risk borrowers will also choose to apply at this interest rate since the net profit of such a borrower in the successful state is

$$135 - 129 = \$6.$$

The total expected profit of Midtown Community Bank, if it lends at an interest rate of 29 percent, is

$$\frac{(0.5 \times 0.9 \times \$129 + 0.5 \times 0.8 \times \$129) \times 1000}{1.05} - \$100{,}000 \qquad [6.3]$$
$$= \$4428.57$$

The expression in (6.3) can be understood as follows. There is a 0.5 probability that the borrower is low risk, in which case the bank gets repaid $129 with probability 0.9. Similarly, there is a 0.5 probability that the borrower is high risk, in which case the bank gets repaid $129 with probability 0.8. This explains the term in the parentheses of the numerator in (6.3). This is multiplied by 1,000 since the bank can make 1,000 such

loans. We discount at the riskless rate of 5 percent since the bank is risk neutral. The initial outlay of $100,000 is finally subtracted to arrive at the bank's expected profit.

Step 2 Since there is unsatisfied loan demand at the 29 percent interest rate—half the loan applicants are turned down—it is natural to ask if Midtown can earn a higher expected profit by increasing the loan interest rate.[3]

Clearly, if you raise the loan interest rate above 29 percent, the low-risk borrowers will not wish to borrow. Since only the high-risk borrowers remain, you might as well raise the loan interest rate all the way up to 34 percent, the maximum you can charge the high-risk borrowers before they too drop out. We refer to 34 percent as a *market-clearing* interest rate since at this level, loan demand equals loan supply.[4]

Midtown Community Bank's total expected profit at this interest rate is

$$\frac{0.8 \times \$134 \times 1000}{1.05} - \$100{,}000 = \$2095.24. \qquad [6.4]$$

Note that (6.4) recognizes that the bank knows that only the high-risk borrowers will apply.

Step 3 It is clear now that the bank earns a greater profit by charging 29 percent and rationing half its loan applicants rather than raising the loan interest rate to a market clearing 34 percent. This illustrates how adverse selection may cause a profit-maximizing bank to ration credit. Raising interest rates in the face of excess demand may drive away the best customers and leave the bank worse off.

1. We could generalize this example to one in which there are numerous imperfectly competitive banks.
2. This assumption is meant to create a strict incentive for the borrower to apply for a bank loan. In its absence, we could have a situation in which the borrower is indifferent between applying and not applying, and then we would need to assume that an application is made in that case.
3. As the ensuing discussion will make clearer, the loan demand curve in this example is downward sloping in the loan interest rate.
4. Since there are 1,000 high-risk loan applicants and each demands a $100 loan, loan demand will be $100,000.

We now turn to an illustration of the *moral hazard effect.*

Example 6.3 Suppose Midtown Community Bank has received a loan application at $t = 0$ from a firm that currently has no assets except for an investment opportunity available one period hence, at $t = 1$. The customer has stipulated that the loan must be made available at $t = 0$ or not at all. The investment outlay required at $t = 1$ is $I_1 = \$100$, of which $55 will come from a bank loan. The firm will make its decision on whether or not to invest at $t = 1$. The firm currently has some securities outstanding. If the investment is made at $t = 1$, it will yield $\$\tilde{y}$ per year perpetually, beginning at $t = 2$. Although $\tilde{y}$ is not known now, it will be known at $t = 1$. There are five possible states of the world at $t = 1$, as shown in Table 6.3 below.

(*Continued*)

TABLE 6.3 Probability Distribution of $\tilde{y}$

State	Probability	$\tilde{y}$
1	0.05	\$15
2	0.05	\$16
3	0.30	\$17
4	0.40	\$18
5	0.20	\$19

Thus, if state 1 is realized at $t = 1$, the project will pay \$15 per year perpetually beginning $t = 2$.

Assume that the riskless rate is 10 percent and the corporate tax rate is zero. Assuming that \$55 of I_1 will be financed with a loan, and the rest will come from the firm's retained earnings, compute Midtown's expected return as a function of the promised loan interest rate. Assume that I_1 is a perpetual loan (a consol) with interest payable at the end of each period, beginning at the end of the first period, that is, at $t = 2$.

Solution The basic idea conveyed by this example is that it does not benefit the bank to keep increasing the loan interest rate because, beyond some point, an increase discourages the borrower from investing when the bank would prefer to proceed with the project. We solve this problem in three steps. First, we provide a framework for linking the bank's *actual* annual interest payment on the loan as a function of the *promised* interest payment. Second, we calculate the interest payment the bank can *expect* to receive each period for different values of the promised loan interest rate. Finally, in Step 3 we conclude that the bank's expected return is maximized at an "interior" loan interest rate, so that if loan demand exceeds loan supply at this rate, the bank will ration credit rather than raise the loan interest rate further.

Step 1 Since at $t = 1$ all uncertainty is resolved, we can view 10 percent as the appropriate discount rate in determining whether or not to undertake the investment at $t = 1$. That is, I_1 will be made at $t = 1$ if and only if $y_s/0.10 \geq I_1$, where $\tilde{y}_s$ is the share of $\tilde{y}$ accruing to the borrower. If the investment is undertaken, then $\tilde{y}_s = \tilde{y}$ – interest on the \$55 loan. Note that the borrower follows this rule because at the time it has to make the investment (at $t = 1$), it already has the money loaned by the bank, and hence treats it as its own retained earnings.

Let r be the actual annual interest payment on the risky bank loan (viewed at $t = 0$, r is a random variable), assuming a perpetual loan with interest payable every period, beginning at $t = 2$. Let r be the *promised* annual interest payment on any debt outstanding at $t = 0$, where r is promised to begin at $t = 2$.

Note that the bank loan is risky only when viewed at $t = 0$. As mentioned earlier, it becomes riskless at $t = 1$. At $t = 1$ then, the value of the bank's loan is the value of a riskless consol bond with an annual coupon equal to the interest payment the bank knows it will receive perpetually, that is, the value of the bank's loan $= \dfrac{\textit{interest payment}}{0.10}$. For example, at $t = 0$ the promised interest payment to the bank may be \$17, but at $t = 0$ we do not know whether this promise can be kept. But suppose at $t = 1$, state 3 is realized. Then, if the firm adopts the project, the promise can be kept for sure, and the $t = 1$ value of the loan is \$17/0.10 = \$170. Alternatively,

if state 2 occurs, the promise will not be kept; the bank will receive only \$16 per year perpetually if the project is adopted. Thus, the time 1 value of the loan is \$16/0.10 = \$160.

Step 2 Now the expected returns to Midtown with different loan interest payments (choice of investment made at t = 1) are given in Table 6.4 below.

TABLE 6.4 Expected Returns to Bank

Promised loan interest $\tilde{r}$	Minimal level of $\tilde{y}$ for investment I_1, to be made by borrowing firm's shareholders	Probability (at t = 0) that investment I_1 will be made	Expected interest payment on bank loan (view at t ≠ 0)
≤\$5	\$15	1.00	$\tilde{r}$
\$6	16	0.95	\$5.70
\$7	17	0.90	6.30
\$8	18	0.60	4.80
\$9	19	0.20	1.80
\$10	20	0.00	0

In this table, the fourth column is obtained by multiplying each promised payment in the first column by the corresponding probability in the third column. The numbers in the third column are obtained by examining the second column and Table 6.3. The smallest possible $\tilde{y}$ in Table 6.3 is \$15, so that the probability of observing a $\tilde{y}$ greater than or equal to \$15 is 1.00. Similarly, from Table 6.3 we see that the probability of obtaining a $\tilde{y}$ at least as great as \$16 is the probability that the state that will occur is either 2, 3, 4 or 5; this probability is 0.95. The rest of the numbers follow similarly.

Step 3 The above table shows that Midtown Community Bank's expected return *peaks* at a promised loan interest of \$7. Note that the present value of the bank loan at $\tilde{r}$ = \$7 is 6.3/0.10 = \$63, which exceeds the loan amount of \$55; hence, Midtown will be willing to lend. Thus, if loan demand exceeds loan supply at that rate, Midtown will be unwilling to extend more credit even if the borrower offers a higher interest rate. Credit rationing occurs here because of moral hazard. However, this moral hazard is a little different from that discussed earlier, wherein the borrower increased the bank's default risk by switching to a risky project from a safe project. Here the borrower prefers not to invest in a project that would have enhanced the bank's expected return; underinvestment is the problem here.

Bank Capital and Credit Rationing

A bank's capital position also may affect its decision to ration credit since different categories of loans have different capital requirements. Consider a bank that has the necessary deposits but would need to raise additional capital to satisfy a loan request.

The additional cost of raising this capital, relative to that of raising money from other sources, will then be a charge against the bank's profit from making the loan. If this additional cost is sufficiently high, the bank may prefer to invest the available deposits in marketable securities rather than in loans. Many allege that this is what happened in 1990–92 and led to a *credit crunch* in the United States despite monetary policy initiatives aimed at reviving the economy.[10]

We have thus far assumed that the bank and the borrower have a one-period relationship. As pointed out earlier, when the bank and the borrower contract with each other over many time periods, it is sometimes possible to reduce informational problems. Indeed, this is one reason to have long-term bank-borrower relationships.

The Spot Lending Decision

We now turn to the bank's lending decision in light of the possibility of credit rationing. To understand this, we should begin by noting that credit analysis, which is an integral part of the lending decision, is not a binary (0 or 1) process whereby the bank either conducts credit analysis or not. It should more appropriately be viewed as a continuum; the bank can perform credit analysis to varying degrees of detail.

The more elaborate the analysis, the more costly it is for the bank. The point to note is that the degree of elaboration is a matter of choice for the bank and represents an important element of the spot lending decision-making process.

The bank must determine its spot lending policy under uncertainty about both the quantity and the quality of loan demand, and within its own capacity constraints. These constraints include limits on screening and monitoring resources. Consequently, the bank may be unable to accommodate more than a predetermined level of aggregate lending without significantly sacrificing loan quality. Loan quality deterioration may imply an unacceptable elevation in the likelihood of ruin for the bank. This means that the first step in lending policy may be for the bank to establish an upper bound, say $\overline{L}$, on the bank's aggregate lending for a given period, say (0, T).[11] Loan applicants arriving after the bank has reached its loan maximum are presumably rejected indiscriminately, and we refer to this phenomenon as *rationing in the large*. Before reaching its loan maximum, the bank does not ration indiscriminately. Rather, it recognizes applicant attributes and rejects only the less desirable. This phenomenon is referred to as *rationing in the small*.[12] The decision to ration an applicant in the small is predicated on the outcome of the bank's credit analysis and its lending prior to the applicant's arrival, as we shall see below.

Consider now a bank that extends \$1 credit to each randomly arriving customer over a fixed planning period (0, T). If a loan applicant arrives at time t, where $0 \leq t \leq T$, the bank conducts credit analysis to estimate the borrower's repayment

10. Thakor (1996) develops a theoretical model that makes precisely this point, and also provides supporting empirical evidence. The model assumes that the additional cost of capital associated with raising capital is exogenously given, and does not provide an endogenous justification for this cost.

11. In the simplest formulation, this capacity constant, $\overline{L}$, can be thought of as a fixed number of dollars, but a more sophisticated formulation might have this capacity a convex and increasing function of the opportunities the bank perceives.

12. Some refer to "rationing in the large" as a borrower being shut out of the bank credit market entirely and "rationing in the small" as loan rejection by an individual bank. Our usage differs.

CHAPTER • 10

The Deposit Contract and Insurance

"As to guaranteeing bank deposits, the minute the government starts to do that, ... the government runs into a probable loss. We do not wish to make the United States government liable for the mistakes and errors of individual banks, and put a premium on unsound banking in the future."

Franklin Delano Roosevelt, in his first press conference as President of the United States

Glossary of Terms

Charter Value: The economic value of a bank to its owners (the shareholders). It can be viewed as the net present value of the profits expected to accrue to the shareholders over the life of the bank.

Null Hypothesis: In statistical-decision theory, when we believe something is true, we formulate the null hypothesis as the alternative to what we believe is true. Thus, when we perform statistical tests using the available data, we *expect* to reject the null hypothesis.

Anticompetitive Restrictions: Restrictions aimed at limiting competition in the banking industry.

Price Elasticity of Demand: A measure of the responsiveness of market demand to changes in price.

Junk Bonds: Very high (default) risk bonds issued by corporations. These bonds have low credit ratings and carry high yields.

Capital Asset Pricing Model: A model describing how risk is priced in the capital market. In particular, it predicts a linear relationship between the expected return on a security and its systematic risk factor (defined as "beta," the ratio of the covariance of the return on the security with that of the market to the variance of the market return).

Introduction

In earlier chapters, we focused on the asset side of the balance sheet for depository institutions. We now shift to the liability side. Although depository institutions have a wide variety of liabilities, in this chapter we will concentrate on different types of deposits and we will turn to capital in Chapter 11.

In the United States, the terms "bank deposits" and "deposit insurance" are almost inseparable. Yet, it is essential to distinguish the issues raised by the deposit contract *per se* from those related to deposit insurance. Thus, we will first discuss the deposit contract without the insurance aspect. We will then discuss "liability management," which is the process of managing the bank's net interest margin, that is the difference between the asset revenues and the liability costs, expressed as a fraction of total assets. This will be followed by an analysis of deposit insurance. Having previously discussed the uninsured deposit contract, we will be able to see how governmental deposit insurance alters the deposit contract, and the behaviors of deposit takers. This, in turn, sets the stage for an analysis of reform proposals in the next chapter.

We doubt that anyone remains to be convinced about the importance of deposit insurance-related issues. There is an almost surreal air about the scandalous 1980s. According to the 1993 *Economic Report of the President*, the S&L industry lost between $100 billion to $160 billion. Commercial banking was shaken to its foundations. Fundamental regulatory reform followed, and a transformation of the financial services industry has occurred as a result.

Many have blamed deposit insurance and greed for the S&L crisis and the widespread banking failures of the 1980s. While this seems to be accepted, it is more difficult to explain why we have deposit insurance, and in particular, why we have the kind of deposit contract that seems to make federal insurance desirable. Discussions of these issues figure prominently in this chapter. Recent events have taught us many valuable lessons. What is unfortunate is that the ongoing crisis was, to a great extent, avoidable, and the regulatory reforms that followed the crisis made sense well before the crisis occurred. As early as 1977, academic publications made the point that federally insured depository institutions had powerful incentives to take asset risk that was excessive from a social welfare standpoint, and that capital regulation, as it existed then, by itself was incapable of controlling these incentives, so that a fundamental reform of regulation was necessary.[1,2]

Some might argue, however, that our historical experience (particularly since the advent of federal deposit insurance, following the Great Depression) did not prepare us for the systemic shocks of the last decade. In the post-1933 period, extremely low bank failure rates[3] made banking a rather unusual industry. So another puzzle is: Why the rash of failures did not occur prior to the 1980s? It turns out that the empirical and theoretical research on which this chapter is based provides valuable insights into the timing of the recent difficulties, and leads us to conclude that, despite our comfortable post-Depression experience, we should have foreseen many of the things that happened.

1. See Kahane (1977) and Merton (1977).
2. See *Capital Issues in Banking* published by the *Association of Reserve City Bankers* (1988).
3. These failure rates were less than 0.3 percent.

The rest of this chapter is organized as follows. In the next section, we discuss the deposit contract. After that, we take up liability management and how it has been affected by interest-rate deregulation and deposit insurance. Then we discuss deposit insurance. We examine the arguments for and against deposit insurance, including the ability of governmentally provided deposit insurance to ward off runs on banks and panics. Issues related to the risk-sensitive pricing of deposit insurance are also examined, as is an analysis of the empirical evidence on the importance of moral hazard in federally insured depository institutions. The empirical evidence also provides insights into the timing of problems with deposit insurance. We then discuss the 1980s deposit insurance debacle in the United States, and developments that have occurred since then.

The Deposit Contract

The Nature of the Deposit Contract

Deposit contracts either have *defined maturities* like certificates of deposit (CDs), or are *withdrawable on demand*. We will focus on demand deposits, the quintessential banking liability. A demand deposit is created when an individual or firm deposits money in an account from which this money can be withdrawn at a moment's notice, that is, on demand.

The demand deposit contract has four important features:

- Its maturity is infinitesimal and can be rolled over indefinitely.
- It is a debt contract.
- It is not traded in a secondary market.
- It is governed by a "sequential service" constraint.

Maturity: The maturity is such that the depositor is promised the ability to withdraw at any time without penalty, that is, the depositor can sell the bank's liability back to the bank at par. Thus, a demand deposit is virtually as liquid as currency. The key difference is that currency carries no default risk, whereas an uninsured bank could default and not be able to fully satisfy withdrawal demands. Indeed, throughout *this* section we will assume that there is no deposit insurance, so that we can focus on the characteristics of the deposit contract itself.

Debt Contract: Because the deposit is a debt contract, the depositor in an uninsured bank confronts the same asset-substitution moral hazard in dealing with the bank as the bank does in dealing with its borrowers (recall Chapters 5 and 6). That is, when a bank creates a deposit, it is simply borrowing from the depositor.

Nontraded Contract: The fact that demand deposits are not traded in a secondary market implies that the depositor's payoff does not depend directly on how information about the bank is processed by other market participants, that is, the depositor does not face market price risk. Unlike a person who plans to sell a traded security in the market at the (random) price prevailing at a future date, a demand depositor knows precisely (in nominal terms) how much she will receive at *any* future point in time when she withdraws from her account, subject to the condition that the bank is solvent.

This last condition is not always satisfied, however. In fact, if things were believed to be going badly for the bank, we would expect the suspicious depositors to rush to the bank to withdraw their deposits. If you arrive late, it is possible that in paying off the earlier depositors the bank will have run out of money by the time you get there. In this case, absent deposit insurance, the maximum amount you can withdraw would be less than you had anticipated. In this sense, your payoff depends on what other depositors believe about the bank, just as it does with any traded debt contract that you liquidate prior to maturity.

The Sequential Service Constraint: This dependence of your payoff on the actions of other depositors occurs because the deposit contract satisfies a sequential service constraint (SSC). Hence, when a depositor seeks to withdraw, the amount the bank pays depends only on what was promised and on his place in the queue of depositors wishing to withdraw. In particular, the depositor's payoff cannot depend on any information that the bank may have about depositors in the queue behind that depositor. Thus, the bank pays depositors on a "first come, first served" basis. To see this, consider a bank that has $5 in equity, and $95 in interest-free deposits acquired from 95 depositors (each of whom deposited $1). The bank's $100 of assets consist of $20 in cash and loans that are currently worth $80 if held to maturity. But if the loans are prematurely liquidated, they are worth only $27.50. Thus, the current (premature) liquidation value of the bank is $47.50. Now imagine that some depositors rush to withdraw their money. Others hear about this and become suspicious about the bank's assets. There is now a full-scale bank run. You are the 48th depositor in a queue of 95 when the bank's doors open in the morning. As the branch manager walks in, she counts the number of people in the queue and sees that every depositor is there to withdraw. Despite this, the SSC dictates that the bank cannot use this information in determining how much the first-in-line depositor should be paid. In this case, the manager is forced to call the outstanding loans, that is, liquidate them to collect $27.50. The first 47 depositors will each receive 41. You will receive $0.50, and all those behind you will go home empty-handed. One might argue that a more equitable approach would have been to give each of the 95 waiting depositors $0.50. But the SSC precludes that.

The nature of the deposit contract is worth examining for two reasons. First, when *all* of the bank's liabilities are uninsured, these features have significant implications for the disciplining of bank management. This suggests that the details of the demand deposit contract are probably not an outcome of chance; they serve a purpose. Second, when deposits are insured, some of these features of the demand deposit contract *encourage* bank runs, thus increasing the liability of the deposit insurer.

The Demand Deposit Contract and Economic Incentives

The Effects of Nontradability and the Debt-Like Nature of Deposits: Consider first that demandable debt is not traded and that it is a debt contract. The analysis in Chapters 5 and 6 implies that the depository institution in this case has an incentive to increase asset risk to the detriment of the depositors. That is, the institution's managers have an incentive to invest in risky loans that transfer wealth from depositors to shareholders. Similarly, depositors face moral hazard in that the institution has an incentive to shirk in monitoring the borrowers to whom it has extended loans. This too

adversely affects the depositors' expected payoff. A third form of moral hazard is fraud. Deposits are essentially "someone else's" money, and managers may be tempted to appropriate some of that money for themselves. While these pathologies have been attributed to federal deposit insurance, they were encountered even prior to the adoption of deposit insurance,[4] and our theory predicts that incentives for managerial fraud exist even with (nontraded) deposits that are uninsured. That the deposit contract is not traded aggravates the moral hazard problem because the discipline imposed by market pricing is absent.

The Effect of Maturity: It turns out, however, that the other two features of the demand deposit contract—its infinitesimal maturity and the SSC—help to attenuate these different types of moral hazard. In developing the intuition below,[5] we first consider the effect of the undefined maturity.

Suppose that there are numerous individuals who demand deposit accounts at a bank. It is natural to expect that some of these depositors are particularly skilled in analyzing the bank's financial health, whereas others are less able. Let us suppose that these skilled depositors keep a watchful eye on the bank's managers because they recognize that moral hazard could diminish their expected payoff. Now, imagine that a few of these vigilant depositors discover that the bank's risky loans are not doing well. Default on many of these loans is likely. Moreover, these depositors discover that the bank has extended numerous loans to close friends of the top managers; this raises suspicion of fraud. What should these informed depositors do? Since they have information that the bank is in peril and may default on its deposit obligations, their best bet is to withdraw their funds as quickly as possible.

When these informed depositors withdraw their funds from the bank, there are two possibilities. One is that the uninformed depositors do not react. In this case, the total outflow of funds from the bank will depend on the size of the deposit holdings of the informed depositors. If their holdings are large enough, the bank will be compelled to attract new deposits. The second possibility is that some or all of the uninformed depositors observe the withdrawals of the informed depositors and decide to follow suit. In this case, there is a bank run. In either case, the bank will need to attract new deposits to replace withdrawals, or liquidate. Liquidation will involve either the calling back of loans, with the associated disruptions in the productive activities of borrowers, or loan sales to other banks. The alternative of attracting new deposits will be difficult, for obvious reasons. Prospective depositors will see the large deposit withdrawals and will be reluctant to entrust their money to the bank. And even if some deposit money flows in, the bank will need to pay higher interest rates on these deposits. Thus, deposit withdrawals by the informed depositors are likely to be costly to the bank. The anticipation of incurring these costs could deter the bank's managers from risky investments, and from shirking on the monitoring of borrowers. It could also reduce the temptation to defraud the depositors.

The Role of the SSC: This argument suggests that the demandable nature of deposits helps to keep bank management on its toes. There is a slight hitch in this disciplining process, however. If a depositor can rely on other depositors to monitor

4. See Calomiris and Kahn (1991).

5. This intuition is based on Calomiris and Kahn (1991), and Calomiris, Kahn, and Krasa (1991). See also Diamond and Rajan (2001), Jacklin (1987, 1989) and Jacklin and Bhattacharya (1988).

the bank, then all that such a depositor has to do is to keep an eye on the informed depositors. There is no need for the "free-riding" depositor to expend personal resources to monitor the bank. This can subvert depositor monitoring. The reason is that *every* depositor may think that others will do the necessary monitoring, and in that case, no one monitors! This is where the SSC comes into play. Because a depositor's expected payoff is greater if he is at the front of the queue than if he is at the rear, he recognizes that by playing a "follow the leader" strategy, his expected payoff is lower than if he monitors himself. This strengthens each individual depositor's incentive to monitor. These ideas are made concrete in the example developed in the box below.

An Illustration of the Incentive Effects of the [Uninsured] Deposit Contract

Example 10.1 Consider a bank that receives a \$1 deposit at $t = 0$ from each of 105 different depositors. It invests \$10 of shareholders' equity in the bank and lends \$110, keeping \$5 as cash reserves. Out of the 105 depositors, there are 30 depositors (called type-D_1 depositors) who are capable of monitoring the bank's management; the remaining depositors (called type-D_2 depositors) keep their money in the bank simply for transactions and safekeeping purposes. The cost of monitoring the bank for an individual type-D_1 depositor is \$0.01 per period.

The bank has two mutually exclusive investment opportunities. Project (or loan) A pays \$200 with probability 0.7 and zero with probability 0.3 at $t = 1$. Project B pays \$150 with probability 0.9 and \$112 with probability 0.1 at $t = 1$. If the bank chooses one of these projects, the probability that the bank will actually end up with that project is 0.9. With probability 0.1, the bank will have inadvertently chosen the other project. Thus, we assume that the bank may make errors in project choice.[1] By monitoring the bank, a type-D_1 depositor can discover the bank's true project choice at some point in time intermediate between $t = 0$ and $t = 1$, say at $t = 1/2$. These depositors can, if they desire, force liquidation of the bank by withdrawing their deposits at $t = 1/2$, and the threat of this liquidation provides a disincentive to the bank to choose the risky project. Note that the bank's projects or loans mature at $t = 1$. If they are liquidated at $t = 1/2$, they are worth only \$25 to the bank. Under the terms of the deposit contract, the bank promises to pay a 12 percent interest (conditional on the bank having the financial capacity to do so) if deposit withdrawal occurs at $t = 1$, and no interest if withdrawal occurs before that. Thus, a depositor is entitled to \$1.12 if she withdraws at $t = 1$, and \$1 if she withdraws at $t = 1/2$. The risk-free discount rate is zero and all agents are risk neutral.

All the type-D_2 depositors plan to withdraw at $t = 1$, but each is subject to a random liquidity-motivated desire to withdraw at $t = 1/2$. To simplify, we will assume that even though no one knows in advance which (type-D_2) depositors will wish to withdraw at $t = 1/2$, the *fraction* of those who will wish to withdraw is known to be 5/75. That is, five type-D_2 depositors will wish to withdraw at $t = 1/2$. Assume that the bank's managers make decisions in the best interests of their shareholders. Compute the equilibrium strategies of the bank and its depositors.

Solution It is useful to summarize the strategies available to the bank and the different types of depositors before we begin to analyze the solution. These are listed below.

TABLE 10.1 Strategies of Participants

Agent	Strategies		
Bank	Choose project A at $t = 0$	Choose project B at $t = 0$	
Type-D_1 depositors	Monitor and decide whether or not to withdraw at $t = 1/2$ based upon result of monitoring	Do not monitor and withdraw at $t = 1$	Do not monitor and withdraw at $t = 1/2$;
Liquidity-motivated Type-D_2 depositors	Withdraw at $t = 1/2$	Withdraw at $t = 1/2$	
Other (Patient) Type-D_2 depositors	Withdraw at $t = 1$	Withdraw at $t = 1$	

We will solve this problem in four steps. First, we analyze the bank's project choice in the case in which the type-D_1 depositors do not monitor and the bank knows that there is no monitoring. We show that the bank chooses project A in this case. Second, we show that our assumption in Step 1 is invalid because it cannot be a Nash equilibrium for no type-D_1 depositors to monitor. Next, we wish to examine if it is a Nash equilibrium for all the type-D_1 depositors to monitor. We do this in two steps. In Step 3, we show that the bank chooses project B if it believes that all the type-D_1 depositors will monitor. Then in Step 4, we examine the strategy of a type-D_1 depositor when he knows that all the other type-D_1 depositors will monitor and the bank has opted for project B. We show that this type-D_1 depositor *will* wish to monitor. This verifies that it is indeed a Nash equilibrium for all the type-D_1 depositors to monitor.

The key assumption in this example is that the bank's project choice cannot be contracted upon because not all depositors can observe it. If this were not the case, there would be no role for depositor monitoring.

Step 1 We will first analyze the outcome in which the type-D_1 depositors do *not* monitor the bank. Given that the bank knows that there is no monitoring, which project will it prefer? If it chooses project A and if this choice is error-free, the expected payoff of its shareholders is

$$\underset{\substack{\uparrow \\ \text{probability} \\ \text{of success}}}{0.7} \times \underset{\substack{\uparrow \\ \text{total} \\ \text{payoff}}}{[200} - \underset{\substack{\uparrow \\ \text{bank's repayment} \\ \text{to its depositors}}}{112]} = \$61.6,$$

and if it chooses project B and this choice is error-free, the expected payoff is

(*Continued*)

$$0.9 \times [150 - 112] + 0.1 \times [112 - 112] = \$3$$

- 0.9: probability of high payoff
- 150: high payoff on project B
- 112: bank's repayment to depositors
- 0.1: probability of low payoff
- 112: low payoff on project B
- 112: repayment to depositors

When project choice is error-prone, the expected payoff of the bank's shareholders when project A is chosen is

$$0.9 \times \$61.6 + 0.1 \times \$34.2 = \$58.86,$$

and their expected payoff when project B is chosen is

$$0.9 \times \$34.2 + 0.1 \times \$61.6 = \$36.94.$$

Thus, if there is no monitoring, the bank will choose project A.

Step 2 The question now is: Can it be a Nash equilibrium for *no* type-D_1 depositor to monitor? This is equivalent to asking whether it is in the best interest of every individual type-D_1 depositor not to monitor when she knows that no other type-D_1 depositors are monitoring. Suppose you are one of those type-D_1 depositors. If you do not monitor, your expected payoff is

$$0.9 \times [0.7 \times \$1.12] + 0.1 \times [\$1.12] = \$0.8176.$$

- 0.9: probability that bank will actually have project A
- $0.7 \times \$1.12$: depositor's expected payoff if bank chooses project A
- 0.1: probability that bank inadvertently chose project B
- $\$1.12$: depositor's payoff when project B is chosen

Now, if you do monitor, and discover that the bank chose project A, what should you do? If you do nothing (that is, you do not withdraw your deposit), your expected payoff at $t = 1$ is

$$0.7 \times \$1.12 - \$0.01 = \$0.774.$$

(0.01: your monitoring cost)

If you withdraw, you know that the bank will be forced to liquidate its asset portfolio since it has only \$5 in cash reserves and there are five type-D_2 depositors who will withdraw at $t = 1/2$ for liquidity purposes. Liquidation will fetch \$25, so that the bank will have a total of \$30 to disburse. You are sure to receive your \$1 at $t = 1/2$. Thus, your payoff will be

$$\$1 - \$0.01 = \$0.99.$$

This means that if you monitor and discover that project A has been chosen, you should demand to withdraw your deposit at $t = 1/2$. On the other hand, if you find that project B was chosen, your payoff is

$$\$1.12 - \$0.01 = \$1.11$$

if you wait until $t = 1$ to withdraw, and it is \$0.99 if you withdraw at $t = 1/2$. Hence, it is better for you to wait until $t = 1$ (remember that your *time value* of money between $t = 1/2$ and $t = 1$ is zero). We can now compute the overall expected payoff to you from monitoring. This payoff is

$$0.9 \times \$0.99 + 0.1 \times \$1.11 = \$1.002.$$

- 0.9: probability that you will discover project A was chosen and will therefore withdraw at $t = 1/2$
- 0.1: probability that you will discover that project B was chosen and will therefore withdraw at $t = 1$

Clearly, this payoff exceeds your payoff if you do not monitor (\$0.8176). This proves that you have an incentive to monitor when others do not, which means that it cannot be a Nash equilibrium for nobody to monitor.

Step 3 Let us now examine if it is a Nash equilibrium for all the type-D_1 depositors to monitor. We begin by noting that if the bank believes that all these depositors will monitor, then it is in the bank's best interest to choose project B. This is verified below.

If the bank chooses project A, then there is only a 0.1 probability that project B will be inadvertently chosen. That is, there is a 0.9 probability that the bank will be liquidated at $t = 1/2$. Thus, the expected payoff of the bank's shareholders from opting for project A is

$$0.1 \times \$34.2 = \$3.42.$$

- 0.1: probability that the bank will not be liquidated
- \$34.2: expected payoff of bank's shareholders if project B is (inadvertently) chosen and bank is not liquidated

If the bank opts for project B, then there is only a 0.1 probability that the bank will be liquidated at $t = 1/2$ (this is the probability that project A will be erroneously picked). Thus, the expected payoff of the bank's shareholders from opting for project B will be

$$0.9 \times \$34.2 = \$30.78.$$

Clearly, the shareholders are better off opting for project B.

(Continued)

Step 4 The next step is to examine the strategy of a type-D_1 depositor when he knows that all the other type-D_1 depositors will monitor and the bank has opted for project B. If you are that depositor and you monitor, your payoff (at $t = 1$) is \$1.12 if you discover at $t = 1/2$ that the bank indeed chose project B. But if you discover that project A was chosen, then you will want to withdraw your deposit. The problem now is a little different from the previous case. You realize that if you discover that project A was chosen, so will the 29 other type-D_1 depositors. When added to the five liquidity-motivated type-D_2 depositors, this means that the line of those who wish to withdraw at $t = 1/2$ will be 35 depositors long. But the bank has only \$30 upon liquidation, and hence can only satisfy the first 30 depositors. Assuming that each person who goes to the bank will have an equal probability of being one of the first 30, the probability is 30/35 that you will be one of the first 30 withdrawers.[2] In this case, your *expected* payoff is only $\frac{30}{35} \times \$1 = \$0.8571$, since you get nothing if you are not one of the first 30 in line. Thus, your overall expected payoff from monitoring is given by

$$0.9 \times \$1.12 + 0.1 \times \$0.8571 - \underset{\substack{\uparrow \\ \text{monitoring cost}}}{\$0.01} = \$1.0837.$$

If you decide not to monitor, then you are behaving like a type-D_2 depositor. Your expected payoff will be \$1.12 if the other type-D_1 depositors discover that project B was chosen (the probability of this is 0.9), and it will be zero if they discover that project A was chosen and decide to liquidate the bank at $t = 1/2$ (the probability of this is 0.1). Hence, your overall expected payoff from not monitoring is

$$0.9 \times \$1.12 = \$1.008.$$

Another possible strategy is for you to behave like a liquidity-motivated type-D_2 depositor and withdraw your deposit at $t = 1/2$ *without* monitoring. In this case, you recognize that there is a 0.9 probability that the other type-D_1 depositors will not withdraw and a 0.1 probability that they will. If the other type-D_1 depositors do not withdraw, there are only six depositors in all (including you) who wish to withdraw at $t = 1/2$. The bank will be forced to liquidate, and you will receive your \$1 for sure. If the other type-D_1 depositors withdraw, the bank will also liquidate, and you will have a 30/35 chance of getting your \$1. Thus, your expected payoff from withdrawing without monitoring is

$$0.9 \times 1 \times \$1 + 0.1 \times \frac{30}{35} \times \$1 = \$0.9086.$$

Comparing the three payoffs (\$1.0837, \$1.008, and \$0.9086), we see that your best strategy is to monitor. Hence, it is a Nash equilibrium for all the type-D_1 depositors to monitor the bank, and for the bank to choose project B.

1. This feature ensures that the type-D_1 depositors do monitor the bank in equilibrium. The reason is that the threat of depositor monitoring will, in equilibrium, cause the bank to choose the project desired by the depositors. If this choice were error-free, depositors would anticipate that the bank will make the desired project choice and therefore perceive no need to monitor. But then the bank, in turn, should anticipate the behavior of the depositors and decide to invest in the project preferred by its own shareholders. And so on and on! The point is that we have a time consistency problem that leads to there being no equilibrium. However, as our solution will make clear this problem can be avoided when the bank's project choice is error-prone.

2. By the SSC, this is the probability that you will receive your \$1.

Although we worked out this numerical example explicitly for the case of asset-substitution moral hazard, the intuition for managerial fraud is similar. In either case, the demandable nature of deposits puts pressure on bank management to not deviate too far from the desires of the depositors, and the SSC lends credibility to the depositors' threat to monitor to ensure "proper" bank behavior by creating a situation in which all vigilant depositors wish to monitor. Thus, these specific features of the deposit contract play an important role in aligning the incentives of the contracting parties in an uninsured bank. This leads naturally to the question of deposit insurance. Before we get to that, however, we discuss liability management in a bank.

Liability Management

We have thus far discussed the economics of the deposit contract. The use of the deposit contract is an integral component of what is called *liability management*.

What Is Liability Management?

Depository institutions pay particular attention to their *net-interest margin* (NIM), which is the difference between the yield on assets and the interest cost of liabilities, expressed as a fraction of total assets. Liability management refers to the institution's strategies for maintaining the continuity and cost effectiveness of funding assets.

There are three main (interrelated) issues in liability management. The first is *diversification*, which refers to choosing among funding sources so as to avoid over-dependence on a particular source. A second choice involves the *mix* of liabilities. Depository institutions raise funds using a variety of deposits, each of which represents a specific contractual form that is a strategic choice. The third choice is about liability *maturity structure*, which determines the bank's interest-rate risk exposure for a given asset maturity structure. We discuss each issue briefly in what follows.

Diversification

Diversifying funding sources reduces liquidity risk (recall Chapter 4). Borrowing and lending in the federal funds market, borrowing at the discount window, dealing with repurchase agreements, and utilizing large CDs, brokered deposits, and Eurodollar deposits are techniques that banks use to diversify. Borrowing in the fed funds market and at the discount window is usually short term; most fed funds transactions are *overnight loans*, although the number of *term fed funds* transactions, with maturities in weeks, has increased recently. For longer-maturity liabilities, banks rely on a variety of deposits. Prominent among these are *negotiable CDs*, called *jumbos*, which are actively traded large-denomination time deposits with market-determined interest rates, a minimum maturity of one week, and denominations exceeding \$100,000. Most negotiable CDs are issued directly to customers, although some large institutions issue them to brokers, who then sell them to other investors. Deposits marketed this way are called *brokered deposits*.

Large banks also use *Eurodollar deposits*, which are time deposits denominated in dollars but held in banks outside the United States, including foreign

branches of U.S. banks. Eurodeposits are created in many ways. Perhaps the simplest way is when an American transfers money on deposit in a U.S. bank to a bank in another country. These deposits remain in dollars. Eurodeposits are subject to the Federal Reserve's cash-asset reserve requirements, and are not protected by U.S. deposit insurance.

Banks also raise funds by using *repurchase agreements* or "repos." A repo is the sale of a marketable security, with the agreement to repurchase it at a specified future date, that is, it is a loan secured by a marketable security. As long as the securities pledged against repos are U.S. government or government agency securities, repos are not subject to reserve requirements. Repos range in maturity from overnight to a month or more. Since repos involve collateral, they are not considered deposits and hence are not covered by deposit insurance.

Banks use a variety of other funding sources, such as subordinated debt as well as securitization and loan sales. Securitization also facilitates diversification of the bank's loan portfolio. Moreover, bank holding companies can issue commercial paper.

Liability Mix

Bank liabilities can be divided into two categories: products and investment instruments.[6] A product entitles the purchaser to a financial claim as well as to some bank services. That is, it is a contract that bundles monetary and possibly nonmonetary payoffs. An example is a checking account on which the bank pays interest and provides transactions services. For corporations, other services include cash management at possibly subsidized prices. Thus, purchasers of product-based deposits, called "customers," receive both explicit and implicit interest, and the demand for such deposits depends both on the explicit interest as well as on the value depositors attach to the bank's services.[7] Because many of these services are demanded by retail depositors, deposits tend to be small (below the *de jure* deposit insurance coverage limit of $100,000 per account). Moreover, customers prefer to have the payoffs on their contracts as insensitive as possible to the fortunes of the intermediary itself. For example, a life insurance policy provides its beneficiaries with a specified cash payment conditional on the death of the insured. That function is less efficiently performed if the contract calls instead for the death benefit to be conditioned on the financial condition of the insurance company as well as on the death of the insured.[8] Consequently, an increase in the policyholder's risk due to a decline in the insurance company's financial condition may require a greater reduction in the insurance premium than would be actuarially fair. It may, therefore, pay for the insurance company to reduce the policyholder's risk as much as possible. In the case of banks, this may explain why product-based deposits are typically fully insured.

6. See Merton (1993).

7. These services are often valued very highly by depositors. Recent empirical evidence has shown that banks enjoy significant economic rents from money-market deposit accounts and NOW (Negotiable Orders of Withdrawal) accounts, both of which are retail deposit accounts. See Hutchison and Pennacchi (1992).

8. Merton (1993) suggests that an Arrow-Debreu economy (see the discussion of market incompleteness in Chapter 1) illustrates this point. A complete set of such securities provides a Pareto-efficient allocation of resources. But efficiency would be lost if the payoffs on such securities were also contingent on the issuer's financial condition [see also Merton (1989)].

Investment instruments, on the other hand, are simply financial claims, similar to the liabilities of nonfinancial firms. The bank provides no transactions or other services to the claimholder, so the design of these contracts involves the same risk-return trade-offs faced by nonfinancial firms. An example of an investment instrument is a brokered CD. Deposit contracts that are investment instruments tend to be purchased by institutions, are relatively large in denominations, and include uninsured deposits. Their prices are usually determined through secondary-market trading.

One of the bank's liability-management choices is the appropriate mix of product-based deposits and investment instruments. Because of the relative insensitivity of their values to the bank's riskiness, product-based deposits do not involve much monitoring of bank management by depositors. Investment instruments, on the other hand, have values that are sensitive to the bank's riskiness, and it pays for the holders of these claims to monitor the bank. The bank is, therefore, subject to greater market discipline with these deposits. From the standpoint of the bank's management, there may be a desire to reduce the bank's reliance on such deposits in order to limit market discipline. Of course, doing so may sacrifice diversification in funding sources, with the attendant liquidity risk that may eventually result in a loss of control for management. The bank's shareholders, on the other hand, would like sufficient reliance on investment instruments to ensure the desired level of market discipline. This suggests a liability-management agency problem between shareholders and managers of banks.

The Duration Structure

Given its asset duration structure, the bank's choice of liability duration structure will determine its interest-rate risk. Given long-duration assets, the bank faces a trade-off in making this choice. On the one hand, choosing a matching long duration on the liability side will minimize interest-rate risk. On the other hand, given the possibility of new information arrival, it may be efficient to choose a shorter duration structure and allow for periodic repricing of deposits. This can reduce the distortions that can arise from private information possessed by the bank at a particular time that may be released to the market later.[9]

Banks often resolve this tension by using derivatives (Chapter 8). The better-managed banks purchase the least expensive assets and liabilities and then use options, futures, and swaps to achieve the desired degree of immunization against interest-rate risk.

Deposit Insurance

The Rationale for Deposit Insurance: A Historical Perspective

The Need for Deposit Insurance: If the demand deposit contract discussed earlier works well in disciplining bank management, why do we need deposit insurance? The reasons are many. Not all make perfect sense in today's environment, but we will get to that later. For now, let us simply note that an uninsured (demand) deposit contract

9. See Flannery (1992).

can be quite disruptive. In a sense, it can lead to *overdisciplining* of banks. This can be seen as follows. In the previous section, we assumed for simplicity that the vigilant depositors could discover the bank's project choice without error. In reality, this discovery is likely to be error-prone. It is then possible that the bank is forced to liquidate assets even when its project choice is congruent with the preferences of depositors. This is socially wasteful *ex post*.

In addition, systematic elements in the risk profiles of the asset portfolios of banks may give risk to a *contagion effect* among banks. That is, when one bank fails, depositors suspect that the failure may be due to systematic risk elements that pervade the asset portfolios of *all* banks in that geographical area, and this may lead to spreading bank runs. Since it often takes a long time for the precise reasons for a bank's failure to become public, the contagion effect may be encountered even when the failure of a particular bank is due to idiosyncratic factors such as poor management. Indeed, this is the rationale for the "too big to fail" doctrine, which leads the government to rescue sufficiently large banks from failure.

Both of these problems are reduced with deposit insurance. When a government agency insures a bank's deposits, it guarantees that the depositors will receive their promised payment, regardless of the bank's financial condition. This makes it unnecessary for depositors to monitor the bank, and it lessens the likelihood of runs on individual banks or on groups of banks.

Historical Background: Federal deposit insurance came into existence in the United States with the enactment of the Banking Act of 1933, and the creation of the Federal Deposit Insurance Corporation (FDIC) to insure bank deposits. The insurance system was extended the following year to S&Ls with the creation of the Federal Savings and Loan Insurance Corporation (FSLIC), which insured S&L shares (deposits). In 1971, deposit insurance was also made available to credit unions.[10] All of this was inspired by the Great Depression and the massive runs on banks that forced President Roosevelt to declare a "banking holiday" in March of 1933. The banking panics of the Great Depression were not new, however. There were as many as seven panics from 1866 to 1934. We will use the term "bank run" (in the singular) to denote a situation in which depositors at a *single* bank wish to exchange their deposits for currency, and the term "banking panic" to "denote a situation in which depositors at many banks wish to exchange their deposits for currency.

Before federal deposit insurance, panics were often addressed by *suspending convertibility* of deposits into cash. Under this approach, the bank was simply closed to depositors who wished to withdraw their money. By giving the banks "breathing room" during which "mass hysteria" had a chance to die down, more information about the financial condition of the bank could be released. Unless this information confirmed the worst fears of depositors, they could be persuaded to refrain from withdrawing their money when the suspension was lifted. Suspension amounted to default on the deposit contract and was a violation of banking law. Nevertheless, five out of the seven panics referred to previously involved suspension of convertibility (those in 1873, 1890, 1893, 1907, and 1914).[11]

10. Legally, a credit union does not accept deposits but issues shares in the credit union to its members. In reality, credit union shares are so similar to deposits that we will not distinguish between them.

11. See Gorton (1988).

Another method that was used during banking panics was the *issuance of clearinghouse loan certificates*. These arose from *Commercial-Bank Clearinghouses* (CBCHs), private-market arrangements among banks that served some of the functions of a central bank. Initially a CBCH was formed to facilitate check clearing.[12] Prior to the formation of the New York CBCH in 1853, for example, commercial banks collected checks by a process of daily exchange and settlement with each other. The clearinghouse centralized the settlement process by permitting exchange to be made with the clearinghouse alone. However, as it evolved, the clearinghouse was able to provide additional information-based services such as *certification* (based on a minimum capital requirement needed to become a member of the clearinghouse) and *monitoring* (based on periodic audits) of its member banks. Members who failed to satisfy CBCH regulations were disciplined with fines or expulsions. This economized on individual monitoring costs that depositors would have had to incur in the absence of a clearinghouse.

One way for a bank to reduce the likelihood of a run is to reduce the depositors' concern about the bank's assets. The clearinghouse loan certificate, first issued during the panic of 1857, was an attempt to do this. A policy committee of the CBCH first authorized the issuance of loan certificates. Whenever a member bank had insufficient cash to satisfy deposit withdrawals, it could apply to the CBCH loan committee for certificates. Borrowing banks were charged interest rates varying from 6 to 7 percent and were required to present acceptable collateral. These certificates, which typically had maturities of 1 to 3 months, could be used by the bank in place of currency. Depositors were willing to accept the loan certificates in exchange for demand deposits because the loan certificates were claims on the CBCH, rather than on the individual bank. Thus, depositors obtained some insurance (diversification benefits) against individual bank failure. This meant that when there was a run on a bank, the bank could either pay off depositors in loan certificates (thereby exchanging claims against its own assets for claims against the CBCH), or it could raise new deposits from depositors who would be sold loan certificates. The bank would then use the proceeds to pay off the older depositors. In this way, the problem of bank-specific risk arising from informational asymmetries was resolved through a private system of coinsurance among banks.

Despite the efforts of the CBCHs to restrain member banks, they could not eliminate *all* moral hazard. Besides, there was the possibility of the CBCH itself being corrupted. Thus, there remained a role for monitoring by depositors. This, in turn, led to occasional runs on banks.

Reasons for Federal Deposit Insurance: Even though private arrangements can diminish the likelihood of bank runs, there are two reasons why they cannot eliminate them. First, even though a private arrangement like the CBCH provides depositors with some diversification, this diversification is limited by the size of the group of member banks. Size limitations may arise from transportation or information costs. Moreover, as the group grows larger, the cost to the CBCH of cheating by an individual bank diminishes, and the CBCH's incentive to monitor its members is weakened. This may be one reason why a large number of new clearinghouses sprang up within a 10-year period following the establishment of the New York CBCH in 1853, rather than a single "mega" clearinghouse emerging. A second weakness of

12. See Gorton and Mullineaux (1987).

private arrangements is that depositors can never be completely sure of the integrity of the arrangement. Thus, there was still some incentive for depositors to monitor the CBCH. In turn, this implies that panics could not be avoided.

The establishment of the Federal Reserve System in 1914 was partly in response to the inadequacy of private arrangements in performing key central bank functions. Nevertheless, the Fed could not prevent the banking panics of the Great Depression, and this eventually led to the establishment of federal deposit insurance. Two of the arguments for federal deposit insurance are discussed below.

(1) Money Supply: the Macroeconomic Argument: At a macroeconomic level, deposit insurance acts as a stabilizer by preventing reductions in the stock of money through bank failures.[13] Since commercial banks are the main providers of the nation's money stock, large-scale uninsured failures of commercial banks would reduce the national money supply. Deposit insurance helps to prevent this in two ways: (a) it replaces deposits that would otherwise be lost, and (b) it discourages banking panics by preserving public confidence.

The reason why deposit insurance has to be *federal* is the credibility of the federal government in its promise to meet all contractual payments. Because of its virtually unlimited authority to raise revenues through taxation, the federal government can meet payout commitments that may be far in excess of the deposit insurance fund. This taxation may be explicit (the government can simply raise taxes) or implicit (the government can print more money to repay depositors, thereby taxing by reducing the real value of each unit of money).

(2) Improving Consumer Welfare: the Microeconomic Argument: We have already noted the incentive of individual depositors to monitor the bank. This results in costly duplication of monitoring. In the numerical illustration of the previous section, the equilibrium involves all 30 vigilant depositors monitoring the bank even though monitoring by just one depositor would suffice. There are two ways in which federal deposit insurance helps to reduce overall monitoring costs. First, because a government agency (the federal insurer) is insuring deposits, the need of insured depositors to monitor is either eliminated (when deposit insurance is complete) or diminished (when deposit insurance is incomplete). Moreover, since the federal insurer must itself monitor banks, even uninsured depositors perceive a much smaller need to monitor. In other words, most of the monitoring burden is shifted from individual depositors to the federal insurer. This eliminates much of the duplicated monitoring encountered with uninsured deposits, *without* any residual monitoring incentives as with the private CBCH arrangement. Second, a federal deposit insurer can be expected to specialize in monitoring insured banks because it must deal with a large number of them. Thus, even apart from reducing duplication, there may be a direct reduction in monitoring costs. For example, in our numerical illustration, instead of monitoring costing 1 cent per audit, it might cost 3/4 cent per audit.

The overall effect of reduced monitoring costs will be to increase the *effective* interest rates on deposits,[14] but this benefit of deposit insurance may be offset by a host of implementation problems that we have yet to address.

13. See Scott and Mayer (1971).

14. To see this, imagine that in the previous numerical illustration, depositors can be assured that the bank will choose project B, and the total monitoring cost to ensure this choice is only 3/4 cent.

Banking Runs and Panics: Theories and the Empirical Evidence

Although the idea that deposit insurance can eliminate bank runs is an old one, research of the last decade has provided a clearer understanding of *why* bank runs and banking panics occur. In light of the recent S&L and banking turmoil, linked by many to federal deposit insurance, alternative arrangements deserve careful consideration. This subsection offers a perspective that should be useful in thinking about these issues.

When informational imperfections interfere with the functioning of a market, governmental intervention may be warranted. An example is Akerlof's lemons problem in the used car market (recall Chapter 1); "lemons laws" protect used car buyers in many states. Another example is the Federal Aviation Authority's regulation of airline safety and the Federal Drug Administration's regulation of the medicinal drug market. In these markets, it is very costly for consumers to let the market provide the necessary disciplining of providers. Similarly, if banking panics disrupt the productive sector of the economy, federal deposit insurance may be warranted if it is effective in reducing the likelihood of panics. The two main theories discussed below explain *how* deposit insurance can prevent runs and panics.

(a) The "Sunspots" Theory of Bank Runs: This theory maintains that bank runs are triggered by completely random events like "sunspots."[15] Suppose that we live in a two-period world with three points in time: $t = 0, 1, 2$. Individuals are risk averse. At $t = 0$, individuals have endowments of wealth that they wish to invest in projects. Each project requires a \$1 investment at $t = 1$, pays off \$R for sure at $t = 2$ if not liquidated earlier, and has positive NPV, that is, each offers a rate of return sufficiently higher than the riskless rate (which is zero) if continued until $t = 2$. Let $R > \$1$. However, if the project is liquidated prematurely at $t = 1$, then there is a loss of productive efficiency and the project pays off only \$1. At $t = 0$, individuals are unsure of their future preferences for the timing of their consumption. At $t = 1$, they receive a "preference shock" and learn whether they are about to die or will live another period. If they are about to die, they want to withdraw the money they have invested and consume it immediately at $t = 1$. If they learn that they will live, then they want to leave their money in the projects and consume \$R at $t = 2$. For the population as a whole, a (random) fraction, f, of individuals are "diers" at $t = 1$ and a fraction, $1 - f$, are "livers."[16]

What would happen without a bank? Well, if you discover at $t = 1$ that you are a dier, you will liquidate your investment and consume \$1. Call the first-period consumption C_1^D, that is, $C_1^D = 1$, and your second-period consumption, $C_2^D = 0$. If you discover that you are a "liver," then you will choose to consume nothing at $t = 1$ (that is, $C_1^L = 0$) and you will consume an amount $C_2^L = R$ at $t = 2$. Thus, the *nonbank outcome* is the pair $\{C_1^D = 1, C_2^D = 0\}$ or the pair $\{C_1^L = 0, C_2^L = R\}$, depending on the individual's type. Is this the best outcome

15. See Noyes (1909) and Gibbons (1968). Bryant (1980) and Diamond and Dybvig (1983) provide contemporary treatments. The discussion below is based on Diamond and Dybvig (1983).

16. The terms "diers" and "livers" are not meant to be taken literally, but merely represent those with preferences for immediate consumption (diers) and for deferred consumption (livers).

from the standpoint of an individual at t = 0? The answer is obviously no! Since you are a *risk-averse* individual, you would like some insurance at t = 0 against a random future shock to your own preference for consumption. This is where a bank can help.

The basic idea is as follows. To provide risk-averse individuals insurance against preference shocks, a bank can arise to promise those withdrawing at t = 1 a little more than $1 and those withdrawing at t = 2 a little less than $R, still ensuring that the promised payoff at t = 2 exceeds that at t = 1. Since R > 1, this is simply a temporal redistribution of the individual's wealth from a state of nature in which wealth is relatively high to one in which it is relatively low, that is, a classic insurance scheme. Compare this to a capital market that also redistributes temporally, but involves no insurance aspect. As long as the bank has a reasonably good idea of how many individuals will withdraw on average at t = 1 (this is similar to insurance companies estimating likely outcomes based on actuarial tables), it can structure the deposit contract in such a way that a known fraction of projects are liquidated at t = 1 to pay off the withdrawers. Note that more projects will need to be liquidated than there are withdrawers because each depositor is promised more than $1 and the liquidation value of each project at t = 1 is $1. Hence, those waiting until t = 2 will receive less than $R. This is a nice arrangement because the t = 2 payoff exceeds the t = 1 payoff, so if a depositor can "afford" to wait until t = 2, he will. Thus, one possible outcome is that only the diers withdraw at t = 1 and all the livers wait until t = 2. All depositors are better off than they would be without a bank because they have received some insurance at t = 0 against unpredictable future changes in their preferences.

The fly in this ointment, however, is that the entire scheme rests delicately on the assumption that none of the livers withdraws at t = 1. But what if a liver believes that others like him might "panic" and withdraw at t = 1? If this belief is justified, it would be foolish for him to be the only patient depositor since the bank will have to liquidate *all* its projects at t = 1 and there will be nothing left to disburse at t = 2. So he will attempt to withdraw at t = 1 as well. In other words, the beliefs of the livers at t = 1 are crucial. If a representative liver believes others will withdraw at t = 1, he will too, and a panic run at t = 1 is a Nash equilibrium. On the other hand, if a representative liver believes others will wait until t = 2, he will too, and this is a Nash equilibrium as well. These beliefs are unrelated to the quality of the bank's assets.

How do you preclude the bad Nash equilibrium? One way is to provide deposit insurance. If the claims of all depositors are insured, then the livers know that they are guaranteed a payoff at t = 2 that is independent of the actions of other depositors. Hence, all livers will withdraw only at t = 2, and there will be no bank run. The example in the box below makes these ideas concrete.

Example 10.2 Suppose there are 100 risk-averse individuals, each with $1 to invest in a project at t = 0. The project will yield $1 if liquidated at t = 1 and $2.25 if liquidated at t = 2. At t = 0, no individual knows what his "type" (denoting his consumption preference) will be at t = 1. If the individual turns out to be a "dier" (type D), then his utility function for consumption will be

$$U_D = \sqrt{C_1^D}.$$

If he turns out to be a "liver" (type L), then his utility function for consumption will be

$$U_L = 0.6\sqrt{C_1^L + C_2^L}.$$

These utility functions capture the idea that the dier benefits from consumption at t = 1 only, and the liver is indifferent between consuming at t = 1 or t = 2 (he gets equal utility from each) so that he will prefer the higher of the two consumptions. It is known at t = 0 that 40 percent of the individuals will end up being diers and 60 percent will be livers at t = 1. Compute the *ex ante* (t = 0) expected utility of each individual if (i) there is no bank and each individual invests in his own projects, and (ii) there is a bank that accepts a $1 deposit from each individual and invests all the proceeds in 100 projects.

Solution We will solve this problem in six steps. First, we calculate each individual's expected utility absent banks. In this scheme, an individual receives $1 if he consumes at t = 1 and $2.25 if he consumes at t = 2. Second, we introduce a bank that is a mutual owned by the 100 depositors. It promises $1.1 to each depositor withdrawing at t = 1 and $2.1 each to those withdrawing at t = 2. Each depositor experiences a higher expected utility at t = 0 with this scheme than in the nonbank case. Third, we show that the intermediated outcome leads to a (good) Nash equilibrium in which all type-D depositors withdraw at t = 1 and all type-L depositors wait until t = 2. Fourth, we show that there is also a bad Nash equilibrium in which all depositors withdraw at t = 1. Fifth, we note that the bank run described in step 4 arises for no particular reason, but that it is possible whenever the existence of the bank makes depositors better off. Finally, in step 6 we show how deposit insurance can eliminate the Nash equilibrium.

Step 1 Consider first the nonintermediated situation. Let us assume, for simplicity, that the diers/livers fractions (0.4 and 0.6) can be viewed as subjective probability assessments of all individuals at t = 0. Then each individual believes that he faces a 0.4 chance of being of type-D at t = 1 and a 0.6 chance of being of type-L. In the nonintermediated case, $[C_1^D = 1, C_2^D = 0]$, and $[C_1^L = 0, C_2^L = R = \$2.25]$. Hence, each individual's expected utility will be

$$\begin{aligned} E(U) &= 0.4 \times \sqrt{1.0} + 0.6 \times 0.6 \times \sqrt{2.25} \\ &= 0.9400. \end{aligned}$$

Step 2 Now consider a bank, owned by its 100 depositors. It provides insurance against depositor preference shocks with a demand deposit offering $C_1^* > \$1$ and $C_2^* < \$R$ (where asterisks denote first- and second-period consumptions in the intermediated case), with the stipulation that C_1^* and C_2^* are mutually exclusive. For example, suppose the bank announces at t = 0 that $C_1^* = \$1.1$. Then, with 40 depositors withdrawing at t = 1, the bank will need to pay out $44, and this requires premature liquidations of 44 projects. The remaining 56 projects will yield a total payoff of 56 × $2.25 = $126 at t = 2. The bank will be able to promise each of the 60 depositors withdrawing at t = 2 an amount $C_2^* = \$126/60 = \$2.1$. The expected utility of a depositor at t = 0 will be

(*Continued*)

$$E^*(U) = 0.4 \times \sqrt{1.1} + 0.6 \times 0.6 \times \sqrt{2.1}$$
$$= 0.9412.$$

Hence, every individual is made better off by the bank that provides *consumption smoothing*.

Step 3 The step-2 outcome is a Nash equilibrium among depositors. Each type-D depositor's Nash equilibrium strategy is to withdraw at t = 1 since that gives him his highest utility (his utility from consumption at t = 2 is zero). If each type-L depositor *takes as given* the Nash equilibrium strategy of the *other* type-L depositors (to wait until t = 2 to withdraw), then no type-L depositor can do better by withdrawing at t = 1. This is because withdrawal at t = 2 gives a type-L a utility of

$$0.6 \times \sqrt{2.1} = 0.8695$$

whereas withdrawal at t = 1 gives a utility of

$$0.6 \times \sqrt{1.1} = 0.6293.$$

Thus, a Nash equilibrium is needed for all type-D depositors to withdraw at t = 1 and all type-L depositors to wait until t = 2.

Step 4 The "good" outcome is not the only Nash equilibrium, however. There is also a "bad" Nash equilibrium with a bank run. To see this, suppose that the representative type-L depositor believes that all the other type-L depositors will withdraw at t = 1 rather than t = 2.[1] What should you, as the "representative" type-L depositor, do?

Suppose you also decide to withdraw at t = 1. The bank will then observe that all 100 depositors wish to withdraw. All 100 projects will have to be liquidated to obtain $100. According to the sequential service constraint, the bank will pay $1.1 each to the first 90 depositors and the remaining $1 to the 91st depositor; the last nine depositors receive nothing. If you wait until t = 2 to withdraw (when all the other depositors withdraw at t = 1), you get nothing. If you rush to the bank at t = 1, then assuming that your position in the queue is decided randomly (with equal probability of being at any position in the queue), you have a 0.9 probability of receiving $1.1, a 0.01 probability of receiving $1, and a 0.09 probability of receiving nothing. Clearly, your optimal strategy is to withdraw at t = 1 too. Thus, it is also a Nash equilibrium for *all* depositors to withdraw t = 1. This equilibrium is a *bank run*.

Step 5 Two points are noteworthy. First, the bank run in step 4 arises for no particular reason. We are not in a position to say which Nash equilibrium will arise. Hence, while we can say that a bank run is a possibility, we cannot say *why*. Second, a simple way for the bank to eliminate this type of run is to stipulate that withdrawers of demand deposits at t = 1 can receive only $1. In this case, the bank does not need to liquidate more projects than there are withdrawers at t = 1, so that a depositor who waits until t = 2 will surely receive $R, Thus, it is optimal for *every* type-L depositor to wait until t = 2, regardless of what the other type-L depositors do. But in this case the bank's demand deposit contract provides no risk sharing and the bank adds no value over the nonintermediated case. Hence, runs are a possibility whenever the bank adds value.[2]

Step 6 Deposit insurance can eliminate the bank run equilibrium *without* trivializing the bank. To see this, imagine that a governmental insurer were to guarantee that any individual withdrawing at t = 1 will receive $1.1 and any individual withdrawing at t = 2 will receive $2.1. Then, only the good Nash equilibrium survives.[3]

1. Do not ask why. This point is to see if this *can* be a Nash equilibrium. That is, conditional on such a belief about the behavior of others, does it pay for the representative type-L depositor to also behave like that?
2. You will note that the bank exists here for a different reason from that in Chapter 3.
3. Suspension of convertibility will work just as well. The bank could announce at t = 0 that only the *first* 40 withdrawers at t = 1 will be paid $1.1 each. Remaining withdrawals can occur only at t = 2. This will do the trick, but only if the fraction of diers is known deterministically at t = 0. If this fraction is random, then the bank will not know *ex ante* when to suspend convertibility. In this case, deposit insurance is necessary to eliminate the bad Nash equilibrium without sacrificing the risk-sharing service banks.

The message of this theory is this: In the absence of deposit insurance, even a perfectly healthy bank faces the threat of a run, given the SSC associated with demand deposits. In other words, runs can result from shifts in the beliefs of individuals, unrelated to the "real" economy or the health of the banking system. Bank runs are simply random manifestations, a *force majeure* triggered even by "sunspots." In French, the term for a bank run is colloquially *sauve qui peut* (every man for himself).

Although some runs reflect sunspot phenomena, it is difficult to verify empirically what precisely triggered a run. Banking panics, on the other hand, have often been triggered by adverse information about banks. We now turn to an informational theory of bank runs.

(b) Adverse Information and Bank Runs: Suppose that we have three types of individuals.[17] We still have the diers (type-D individuals) who must consume at the end of the first period and represent a fraction, f, of all individuals. But among the livers, (type-L individuals), we now have a fraction who receives information about the terminal (t = 2) value of the bank's assets. In the previous theory, we assumed that this value, $R, was nonrandom and known to everyone. Assume now that $\tilde{R}$ is a random variable with a commonly known expected value, R. Let $\tilde{R} = H > 0$ with probability p and $\tilde{R} = 0$ with probability 1 − p. Thus, at t = 0, no individual knows either the t = 2 value of $\tilde{R}$ or what his type (D or L) will be at t = 1. However, at t = 1, each individual discovers whether he is a D or L, and some fraction, q, of the Ls also comes to know the value $\tilde{R}$ will take at t = 2. Nobody knows how many individuals of each type there are at t = 1 (that is, both the fraction f and the fraction q are random).

17. This discussion is based on Chari and Jagannathan (1988).

The choice problem faced by the Ds and the informed Ls at t = 1 is straightforward. All the Ds will line up to withdraw their deposits. If the informed Ls learn that R = H, then it is better for them to defer withdrawal until t = 2, thereby avoiding premature project liquidation. But if the informed Ls learn that R = 0, then it pays for them to withdraw whatever they can at t = 1.

Consider now the choice problem of the uninformed Ls. They can withdraw at t = 1 or wait until t = 2. Their decision will be based on their assessment of the t = 2 value of the bank's assets. Although they cannot directly observe this value, they can infer it by observing the length of the withdrawal queue at t = 1.[18] In drawing this inference, they realize that some people are in the withdrawal queue at t = 1 because they have discovered that they are Ds. But they do not know *how many* such individuals there are. This means that when they observe the length of the withdrawal queue at t = 1, they are unsure whether all are Ds or whether some are informed Ls.

It is true, however, that the longer the queue the more likely it is that it contains some informed Ls with adverse information about the bank. If the uninformed Ls knew for sure that the queue contained informed Ls, they would withdraw their money at t = 1, and if they knew for sure that it contained only Ds, they would defer withdrawal until t = 2. But when they cannot be sure, they use the queue length as a *noisy* signal of the information possessed by the informed Ls. Thus, they withdraw their deposits at t = 1 if the queue is sufficiently long, and they defer withdrawal until t = 2 if the queue is shorter.

Defining a bank run as a situation in which uninformed Ls withdraw at t = 1, we see that a bank run is more likely when some depositors receive adverse information about the bank. The reason is that as the informed Ls line up to withdraw their funds, they increase the queue length. This induces the uninformed Ls also to seek withdrawal of their deposits. Thus, a bank run results from depositors attempting to detect the bank's condition from the length of the withdrawal queue. However, since their learning is "noisy" (they occasionally confuse liquidity-motivated withdrawals with informed withdrawals), they make both type-I and type-II errors.[19] That is, they sometimes do not run the bank when they should (when the queue is relatively short but consists of informed Ls: a type-II error if the null hypothesis is that the bank is healthy); and they sometimes run the bank when they should not (when the queue length is relatively long but consists only of Ds: a type-I error). Because runs can sometimes occur when they should not, deposit insurance may improve welfare by eliminating the possibility that uninformed Ls will erroneously withdraw.

(c) The Empirical Evidence on Panics: Strictly speaking, neither of the two theories of bank runs discussed just above explains *panics*. According to the sunspots theory, a bank run is a completely random event, so there is no reason for a run to precipitate a panic, although a panic could come about by pure chance. According to the adverse information theory, a run is caused by information *specific* to a bank. Once again, there is no reason for a run to be contagious. These then are theories of bank runs and not banking panics.

The adverse information theory, however, can be adapted to provide an explanation for banking panics. Suppose there is information about some event that is relevant to the fortunes of all banks. That is, there is a systematic risk element that affects all banks.

18. This inference will usually be noisy. Formally, the inference may be made using Bayes' rule (see Chapter 1).

19. A type-I error in statistics is when the decision maker rejects the (null) hypothesis although it is true and a type-II error is when he accepts (or more appropriately, fails to reject) the null hypothesis although it is false.

Unlike the standard Capital Asset Pricing Model, however, assume that the systematic risk is *not* commonly known. Individuals may then attempt to infer something about systematic risk from their observations of presumably related events. For example, the failure of a large bank may cause depositors to believe that general economic conditions have deteriorated, and this may lead to a panic. The intuition is similar to that of the adverse information theory. According to that theory, depositors infer something about their bank from the behavior of fellow depositors. Here, depositors at one bank infer something about their bank from the behavior of depositors in *other* banks.

An example of an event that may reveal adverse systematic information is a recession, or a bank run during a recession. During the period from 1873 to 1914, every major business cycle downturn was accompanied by a banking panic.

Empirical evidence supports this version of the adverse information hypothesis. If banking panics are indeed systematic events, then there must be a change in the risk perceptions of individuals prior to a panic, and this, in turn, must cause a change in the deposit/currency ratio. That is, the perceived risk variable must achieve some critical value at the panic date. Also, the movements in the risk predictors and in perceived risk should occur at panic dates and not at other dates. If such movements occurred at other dates, then there should have been panics at those dates.

An empirical examination of panics in the pre-Federal Reserve era provides insight into the relationship between changes in risk perceptions and banking panics.[20] To serve as a proxy for perceived risk, empiricists use unanticipated changes in the liabilities of failed businesses.[21] This is reasonable since the fortunes of nonfinancial firms affect the fortunes of banks. As Table 10.2 shows, panic dates correspond to the timing of the largest values of the liabilities shocks. Panics also follow the business cycle peak by several months.

The study also indicates that the percentage change in the currency/deposit ratio is significantly correlated with the perceived risk measure. Thus, the data for the pre-Fed period support the notion of a threshold value of perceived risk that triggers panics. More recent research indicates that the banking panics during 1890–1909 were triggered by net movements of deposits away from the money-center banks and low levels of excess reserves. Changes in stock market values had little effect.[22]

The formation of the Federal Reserve System in 1914 and the initiation of deposit insurance in 1934 had a significant influence on the timing of panics. In the period from 1914 to 1933, we see from Table 10.3 that changes in the perceived risk measure were large enough in at least one instance (June 1920) to cause panics during the pre-Fed period, but resulted in no panics in the post-Fed period.

The introduction of deposit insurance again significantly changed depositor behavior. In the period from 1935 to 1972, until after deposit insurance was introduced, there were several instances of large failed business liabilities shocks, none of which resulted in panics. Thus, deposit insurance appears to have served its purpose.

Deposit Insurance Pricing and Moral Hazard

Until the 1980s, the pricing of federal deposit insurance was largely *risk insensitive*. That is, each bank was charged an insurance premium that depended only on its

20. The evidence reported here is from Gorton (1988).

21. In Gorton's (1988) empirical study, this variable was measured by the residuals ("error terms") from an estimated time-series model.

22. See McDill and Sheehan (2006).

TABLE 10.2 The Relationship Between the Timing of the Largest Unanticipated Changes in the Liabilities of Failed Businesses and the Timing of Banking Panics in the National Banking Era

NBER Chronology Peak-Trough (Business Cycle)	Timing of Largest Value of Unanticipated Changes in Liabilities of Failed Businesses	Panic Date
Oct. 1873–Mar. 1879	Dec. 1873	Dec. 1873
Mar. 1882–May 1885	June 1884	June 1884
Mar. 1887–Apr. 1888	Nov. 1887	No panic
July 1890–May 1891	Dec. 1890	Dec. 1890
Jan. 1893–June 1894	July 1893	July 1893
Dec. 1895–June 1897	Oct. 1896	Oct. 1896
June 1899–Dec. 1900		No panic
Sep. 1902–Aug. 1904		No panic
May 1907–June 1908	Feb. 1908	Dec. 1907
Jan. 1910–Jan. 1912	Mar. 1910	No panic
Jan. 1913–Dec. 1914	Mar. 1914	Sep. 1914

Source: Gorton, Gary, "Banking Panics and Business Cycles," *Oxford Economic Papers* 40, 1988, 751–781.

TABLE 10.3 The Relationship Between the Timing of the Largest Unanticipated Changes in the Liabilities of Failed Businesses and the Timing of Banking Panics in the Federal Reserve Era

Peak-Trough (Business Cycle)	Timing of Largest Value of Unanticipated Changes in Liabilities of Failed Businesses	Panic Date
Aug. 1918–Mar. 1919	Nov. 1918	No panic
Jan. 1920–July 1921	June 1920	No panic
May 1923–July 1924	Nov. 1923	No panic
Oct. 1926–Nov. 1927	Apr. 1927	No panic
Aug. 1929–Mar. 1933	Dec. 1929	Oct. 1930
		Mar. 1931
		Jan. 1933

"The change in perceived risk in June 1920 was large enough to have caused a panic in the pre-Fed Era."

Source: Gorton, Gary, "Banking Panics and Business Cycles," *Oxford Economic Papers* 40, 1988, 751–781.

volume of deposits, and not on its riskiness. Many have charged that this heightened incentives for insured depository institutions to take excessive levels of risk. Note that institutions like banks can increase risk in a variety of ways. However, for the purposes of this discussion, we will focus on the bank's incentive to invest in assets with high default risk. Although deposit insurance premiums are now risk sensitive, only a limited number of risk categories are used and at best, the premiums are only crudely related to risk for most banks. In this section, we will show how the imperfectly risk-sensitive structure of deposit insurance pricing also creates incentives for excessive risk-taking by banks.[23]

23. The discussion here is based on Merton (1977).

Deposit Insurance as an Option: Consider an insured bank (both principal and interest on deposits are insured) that has raised deposits requiring the bank to repay \$B at the end of the period. Let \$V be the total value of the bank's assets at the end of period. Now, if $V \geq B$, then the depositors receive \$B from the bank and the bank's shareholders receive $\$(V - B)$. If $V < B$, then the bank fails. Its shareholders receive nothing, whereas the deposit insurer takes possession of the bank's assets and pays out \$B to the depositors. The net *loss* to the deposit insurer in this case is $\$(B - V)$. Thus, the end-of-period payoffs to the different parties can be written as

Shareholders:	Max[0, V − B]
Depositors:	B
Deposit Insurer:	Min[0, V − B], which is either zero (when V > B) or negative (when V < B).

The effect of deposit insurance is to create an additional cash inflow to the firm of −Min[0, V − B] dollars. But −Min[0, V − B] can also be written as Max[0, B − V]. Hence, if G(T) is the value to the firm of the deposit insurance guarantee when the length of time remaining to maturity of the deposits is T, then on the date of maturity,

$$G(0) = \text{Max}[0, B - V]. \tag{10.1}$$

You will recall now from our discussions of options in Chapters 1 and 8 that the payoff structure in (10.1) is identical to that of a *put option* at expiration. To see this, imagine that V is the (random) value of the underlying security on which the option is written, and B is the exercise (or strike) price. Then, as the owner of the put, you will exercise your option to *sell* the security to the option writer at \$B if the value of the security, V, is less than B. In this case, your gain from exercising the option will be $\$(B - V)$. On the other hand, if $B < V$, then you will let the option expire unexercised, and your gain will be zero.

The Cost of the Option: In other words, when the FDIC insures a bank's deposits, it is writing a put option in favor of the bank. The cost to the FDIC of providing this insurance is simply the value of the put option. We can calculate this value using the option pricing formula developed by Black and Scholes (1973):

$$G(T) = Be^{-rT}\Phi(x_2) - V\Phi(x_1) \tag{10.2}$$

where

$$x_1 \equiv \frac{\log(B/V) - \left[r + \frac{\sigma^2}{2}\right]T}{\sigma\sqrt{T}}$$

$$x_2 \equiv x_1 + \sigma\sqrt{T}.$$

Here r is the instantaneous risk-free interest rate, $\Phi(\bullet)$ is the standard normal cumulative distribution function, V is the current value of the bank's assets, and σ^2 is the variance rate per unit time of the logarithmic changes in the value of the assets. It is assumed that all the Black-Scholes assumptions are satisfied.

The Cost Per Dollar of Deposits: We can also compute the appropriate deposit insurance premium per dollar of deposits. If depositors are promised a repayment of \$B at a time, T, in the future, then the current value of these (riskless) deposits will be

$$D = Be^{-rT}. \tag{10.3}$$

Let $g = G(T)/D$ be the cost (to the FDIC) of the deposit insurance guarantee per dollar of insured deposits. Then, using (9.2) and (9.3) we can write

$$g(d,\tau) = \Phi(h_2) - \frac{1}{d}\Phi(h_1) \tag{10.4}$$

$$\text{where } h_1 \equiv \frac{\left[\log - \frac{\tau}{2}\right]}{\sqrt{\tau}} \tag{10.5}$$

$$h_2 \equiv h_1 + \sqrt{\tau}. \tag{10.6}$$

Here $d \equiv D/V$ is the current deposit-to-asset value ratio for the bank, and $\tau \equiv \sigma^2 T$ is the variance of the logarithmic change in the value of the assets during the term of the deposits.

Properties of a Risk-Sensitive Deposit Insurance Pricing Scheme: A few points are worth noting. First, an increase in the deposit-to-asset value ratio causes an increase in the cost per dollar of deposit insurance to the FDIC, that is,

$$\partial g/\partial d = \Phi(h_1)/d^2 > 0.$$

Similarly, as τ increases, so does the cost of deposit insurance, that is,

$$\partial g/\partial \tau = \Phi'(h_1)/2d\sqrt{T} > 0.$$

Here the prime denotes a derivative; hence, $\Phi'(h_1)$ is the standard normal density function at h_1. This is a well-known property of options; their value increases with the volatility of the underlying security. Hence, the FDIC should charge a higher deposit insurance premium for banks with lower capital-to-total-assets ratios and higher volatility in the value of total assets. Alternatively, in a regime in which the FDIC charges each bank a fixed premium per dollar of insured deposits, rather than g (which is a function of d and τ), banks with higher capital ratios and lower asset risks *subsidize* those with lower capital ratios and higher asset risks, assuming that the FDIC breaks even on average.

The Option Feature and Moral Hazard: These observations also highlight the moral hazard inherent in deposit insurance. Since g is the *value* to the bank of deposit insurance per dollar of insured deposits, a bank can increase this value by reducing its capital and increasing its asset volatility. To the extent that the premium charged is insensitive to these initiatives of the bank, a *shareholder-wealth-maximizing* bank has an incentive to increase financial leverage and asset volatility. *Figure 10.1* illustrates this incentive graphically.

In Figure 10.1, the curve AB is the total expected return on the bank's assets, net of bankruptcy costs.[24] This curve peaks at σ^*. The expected return to depositors, as represented by the straight line CD, remains constant because we assume that deposits are completely insured. The total expected return to depositors and the FDIC is equal to the deposit yield plus the deposit insurance premium minus the expected bankruptcy costs. This total *expected* return, represented by the curve BF, is constant for $\sigma < \sigma_f$ (some threshold variability) because the probability of bankruptcy is zero in this range. Then, as the probability of bankruptcy rises, the total expected return to the FDIC and the depositors declines. Since the depositors are completely protected and the deposit insurance premium is insensitive to σ, it is the expected return to the FDIC that falls very rapidly as σ arises. Consequently, even though the total expected return on the bank's assets is falling as σ increases beyond σ^*, the expected return to the bank's shareholders is increasing in this range. In fact, the shareholders' expected return peaks at $\sigma_m > \sigma^*$.

The optimal level of risk (as represented by σ) depends on the decision maker's objective. If the objective is to minimize the liability of the deposit insurer, then the optimal risk choice is $\sigma = \sigma_f$. If the objective is to maximize the bank's total expected return on assets, then the optimal risk choice is $\sigma = \sigma^*$. But if decisions are made to maximize the wealth of shareholders, the optimal risk choice is $\sigma = \sigma_m$. Thus, if the "socially desired" risk choice is σ^*, the bank will take more risk than the social optimum by choosing σ_m. This is the moral hazard of deposit insurance.

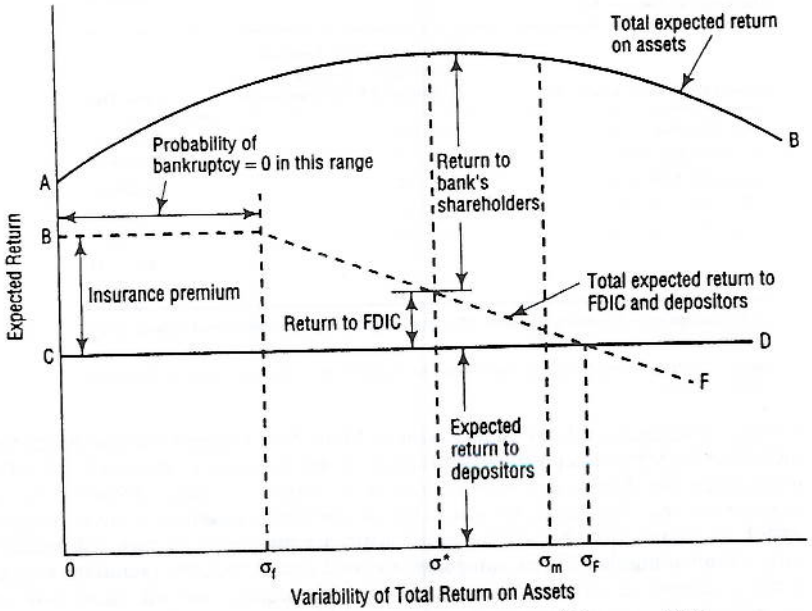


FIGURE 10.1 The Relationship Between Expected Return and Risk

24. Figure 10.1 is based on Keeton (1984), who provides a similar figure (see p. 32 in that paper).

Why the Concern With Moral Hazard in Banking?: You will recall from Chapters 5 and 6 that a similar moral hazard exists for nonfinancial firms that borrow from banks. However, with nonfinancial firms, the costs of this moral hazard are borne *ex post* by private lenders, who pass along these costs *ex ante* (through the pricing mechanism) to the borrower. Thus, the moral hazard gets priced among the contracting parties in equilibrium. In the case of banks and other federally insured depository institutions, however, these costs are borne *ex post* by the FDIC, and hence, eventually by the taxpayers. Of course, if the FDIC breaks even in aggregate, then these costs are passed along *ex ante* to the banking industry as a whole, and there is simply a redistribution of wealth across banks. That is, less risky banks end up subsidizing their riskier counterparts, with no direct wealth consequences for the taxpayers.

This analysis, as well as our earlier discussion of the similarity between a deposit insurance guarantee and a put option, indicates a role for safety regulation in banking. Given deposit insurance, banks have a propensity to lower capital and increase risk. Capital requirements and asset portfolio restrictions seek to address these distorted incentives arising from deposit insurance. However, the implementation of these regulatory devices has not always been effective.

In the box below, we provide an illustration of the effect of moral hazard in the context of the put option pricing formula.

Example 10.3 Consider a bank with federally insured deposits maturing in one year. Imagine that the bank's asset value changes monthly and you have been provided the following data on asset values for the past seven months (you may assume that the probability distribution of asset value changes remains stationary through time).

Month	Bank Asset Value (in millions of dollars)
1	100
2	101
3	99
4	102
5	100
6	98
7	97.605074

Suppose the bank's current deposit-to-asset value ratio is 0.95. Compute the value to the bank of the deposit insurance guarantee per dollar of insured deposits. Also compute the value of this guarantee for a higher deposit-to-total-asset-value ratio (of your choice), holding fixed the variance of asset value changes, and the value of this guarantee for a higher variance, holding fixed the deposit-to-total-asset-value ratio.

Solution We solve this problem in three steps. First, we will compute τ, the variance of bank asset values. Second, we compute h_1 and h_2 using the value of τ obtained in the previous step. Finally, in step 3 we calculate the cost of deposit insurance per dollar of insured deposits.

Step 1 To compute τ, we define V_t as the asset value in month t and V_{t-1} as the asset value in month $t-1$. Thus, when we write the asset value in month 2, for example, we will write V_2, and when we write the ratio V_t/V_{t-1} in month 2, we will write V_2/V_1. We can construct the following table.

TABLE 10.4 Calculation of Asset Value Variance

A Month	B Asset Value V_t	C V_t/V_{t-1}	D $\log(V_t/V_{t-1})$	E D-Sample Mean	F $(E)^2$
1	100	—	—	—	—
2	101	1.01	0.00995	0.013988	0.0001957
3	99	0.9802	−0.02000	−0.015962	0.0002548
4	102	1.0303	0.02985	0.033888	0.0011484
5	100	0.9804	−0.01979	−0.015752	0.0002481
6	98	0.9800	−0.0202	−0.016162	0.0002612
7	97.605074	0.9959701	−0.004038	0	0

In this table we compute the "sample mean" by adding up the entries in column D and dividing by 6 to obtain −0.004038. Column E is then obtained by subtracting the sample mean from each entry in column D. Column F is merely each entry in column E squared. Now,

$$\sigma^2 = \frac{\text{sum of all entries in column F}}{5}$$

$$= \frac{0.0021082}{5} = 0.0004216.$$

Note that we divide by 5 because we lose one degree of freedom in computing the variance. Now, $\tau = \sigma^2 T = 0.0004216 \times 12 = 0.005$ approximately. Note that $T = 12$ since the deposit maturity is 1 year and asset values change monthly.

Step 2 Next, we compute h_1 using (10.5) as

$$h_1 = \frac{\log(0.95) - (0.005/2)}{\sqrt{0.005}}$$

$$= -0.76076$$

and h_2 using (10.6) is

$$h_2 = -0.76076 + \sqrt{0.005} = -0.69005.$$

Step 3 Using (10.4), we can now compute g as

$$g = \Phi(-0.69005) - \frac{1}{0.95}\Phi(-0.76076) \cong 0.0099.$$

Thus, the value to the bank of having the deposit insurance guarantee is roughly 99 cents per \$100 of insured deposits. This is much higher than the current premium of approximately 25 cents per \$100 of insured deposits. In Table 10.5, we present calculations for a variety of deposit-to-asset value ratios and values of τ. Note that if we increase d to 1 and hold τ fixed at 0.005, the value of g rises to \$2.82 per \$100 of insured deposits. This illustrates the bank's incentive for leverage emanating from deposit insurance. Similarly, if we hold d fixed at 0.95 and increase τ to 0.006, the

(Continued)

value of g rises to $1.209 per $100 of insured deposits. This illustrates the bank's incentive to take on more risky assets.

TABLE 10.5 Cost of Deposit Insurance per Dollar of Insured Deposits

Cost of Deposit of Insurance (g)	Deposit-to-Asset Value Ratio	Variance (τ)
0.00055	0.85	0.00600
0.00040	0.85	0.00550
0.00028	0.85	0.00500
0.00018	0.85	0.00450
0.00011	0.85	0.00400
0.00326	0.90	0.00600
0.00274	0.90	0.00550
0.00223	0.90	0.00500
0.00176	0.90	0.00450
0.00132	0.90	0.00400
0.00093	0.90	0.00350
0.00060	0.90	0.00300
0.00015	0.90	0.00200
0.01209	0.95	0.00600
0.01102	0.95	0.00550
0.00992	0.95	0.00500
0.00880	0.95	0.00450
0.00765	0.95	0.00400
0.00647	0.95	0.00350
0.00528	0.95	0.00300
0.00287	0.95	0.00200
0.00172	0.95	0.00150
0.00072	0.95	0.00100
0.00033	0.95	0.00075
0.03089	1.00	0.00600
0.02958	1.00	0.00550
0.02820	1.00	0.00500
0.02676	1.00	0.00450
0.02523	1.00	0.00400
0.02360	1.00	0.00350
0.02185	1.00	0.00300
0.01784	1.00	0.00200
0.01545	1.00	0.00150
0.01262	1.00	0.00100
0.01093	1.00	0.00075
0.00892	1.00	0.00050
0.00631	1.00	0.00025
0.00564	1.00	0.00020
0.00489	1.00	0.00015
0.00399	1.00	0.00010
0.00282	1.00	0.00005
0.00126	1.00	0.00001

Source: Merton, Robert C., "The Cost of Deposit Insurance and Loan Guarantees," *Journal of Banking and Finance* 1, June 1977, 10.

The option pricing approach indicates factors that must be considered in setting the deposit insurance premium. The premium per dollar of insured deposits must be sensitive to the volatility of the bank's assets and to its deposit-to-total-asset ratio. If not, the bank will have an incentive to reduce its capital and increase its asset risk in the interests of its shareholders. The option pricing approach is not meant to be taken literally as a precise way to set the deposit insurance premium, since many of the standard Black-Scholes assumptions are not satisfied.[25] For example, the asset values of banks often exhibit jumps rather than following a continuous path through time as assumed by Black-Scholes. In any case, the numerical values in Table 10.5 suggest the magnitude of the gains to banks from exploiting risk-insensitive deposit insurance pricing.

Empirical Evidence on Moral Hazard

Apart from the anecdotal evidence on moral hazard, there is now substantial scientific evidence to support the theories we have reviewed. Federal deposit insurance has been in existence for banks since 1934, but the more visible problems were encountered only during 1970–90. This suggests that there must have been *countervailing forces* in the past that diminished the risk-taking propensity created by deposit insurance. The empirical evidence we discuss here sheds some light on these forces.

(a) Some Evidence on the Effect of Federal Deposit Insurance on Risk-Taking: The Case of Credit Unions: Federal deposit insurance was extended to credit unions in 1971 when the National Credit Union Administration (NCUA) was formed, and the coverage limits are currently the same as for banks and thrifts.

Credit unions: (i) make loans to their own members, (ii) make loans to other credit unions, and (iii) engage in loan participations with other credit unions. A credit union's asset portfolio consists primarily of: (i) secured loans for the purchase of consumer durables, and (ii) investments in low-risk assets such as government bonds, loans to other credit unions, and deposits with commercial banks.

A credit union can increase its risk by decreasing its capital cushion and by increasing the fraction of its total assets invested in high-risk assets. An empirical examination provided support for this hypothesis.[26] *Figure 10.2* is a graph illustrating the behavior of capital ratios (defined as capital divided by total assets) for federal credit unions over the 1949–1992 period. In 1970, just prior to the adoption of federal deposit insurance, the capital ratio was about the same as in 1949. There was a slight decline during the transition period from the preinsurance regime to the insurance regime. The sharpest decline occurred during the insurance period. This is consistent with the prediction of a moral hazard associated with deposit insurance.

(b) The Relationship Between Market Power in Banking and Moral Hazard: As mentioned earlier, a major puzzle is why the deposit insurance system in the United States worked so well for so many years despite the risk-taking incentives provided by federal deposit insurance and why problems surfaced only recently. One explanation is that a bank's risk-taking propensity depends on the value of its charter. The higher the charter value—the capitalized value of its future cash flows—the weaker is the

25. In addition, it may be difficult to ensure that the deposit insurer measures bank risk without error. See Flannery (1991) for a discussion of the implications.

26. See Clair (1984).

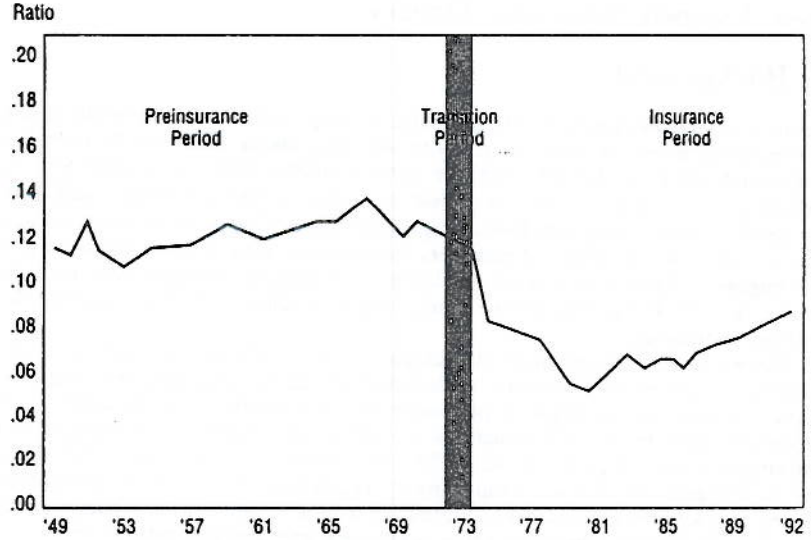


FIGURE 10.2 Capital Ratio for Federal Credit Unions
Source: National Credit Union Administration.

bank's incentive to take risk. This is because higher risk implies a higher likelihood of insolvency, in which case the insurer takes possession of the bank, and the charter is lost. Thus, the higher the value of this charter, the greater is the bankruptcy cost for the bank. In the past, various anticompetitive restrictions gave banks market power that enhanced the value of charters. The loss to the bank from losing its charter in the event of bankruptcy provided a counterbalance to the incentive for excessive risk-taking due to fixed-rate deposit insurance.[27] The deregulation that took place in the 1980s increased banking competition but lowered the value of bank charters. Greater risk-taking was predictable.

Evidence supports this theoretical prediction. *Figure 10.3* is a graph of the time series behavior of the average capital/total assets ratio of the 25 largest bank holding companies in the United States from 1952 to 1986. The decline in this ratio is significant.

A *direct* test of the relationship between risk-taking and charter value would need to have some measure of the capitalized value of future rents, or market power. One such measure is "Tobin's q," which is approximated as the ratio of the market value of assets (market value of common equity plus the book value of liabilities) to the book value of assets. The higher the q ratio, the larger is the charter value, relative to the book value of its assets. Since bank risk-taking is also not directly observable, a proxy is needed. A reasonable proxy is the interest cost on large uninsured CDs. The

27. For the theory, see Chan, Greenbaum, and Thakor (1992). The empirical evidence discussed below is from Keeley (1990).

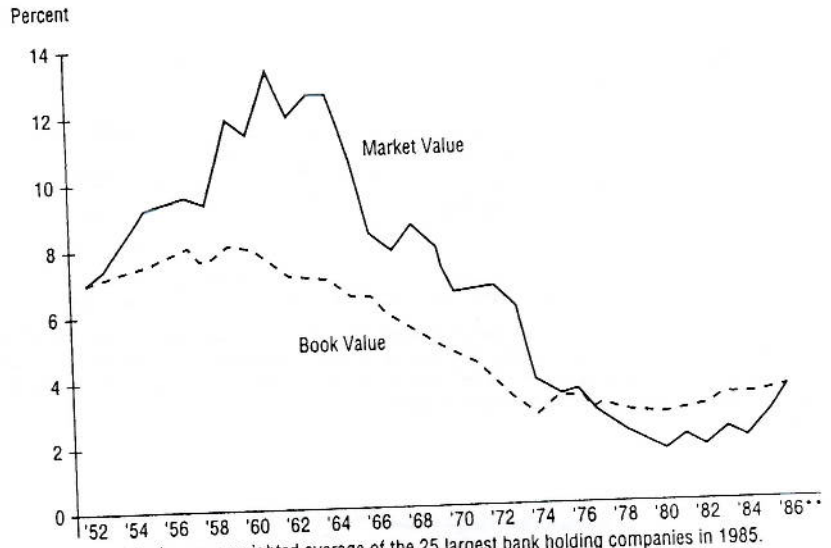


FIGURE 10.3 Capital-to-Asset Ratios, Market and Book Values
Source: Keeley, Michael C., "Deposit Insurance, Risk, and Market Power in Banking," *American Economic Review* 80, December 1990, 1183–1200.

holders of such CDs should be sensitive to the bank's riskiness and demand higher interest rates from riskier banks. The evidence is quite compelling. Each 1 percent increase in the q ratio results in a 16 to 18 basis point reduction in the average CD cost. Moreover, this relationship is statistically significant. Thus, bank risk-taking appears to have increased substantially in the 1980s owing to deregulation that diminished bank charter values.

To provide a comparison with more recent data, we have provided in *Figure 10.4* the capital ratios in book and market value terms for the 25 largest bank holding companies from 1959–2005. The effect of the capital regulation that began with the Basel I Accord is evident, as capital ratios exhibit an upward drift beginning in the late 1980s.

We also show the capital ratios during 1992–2005 in *Figure 10.5*. This figure shows that capital ratios have remained relatively flat during this time in book value terms, and above the minimum requirement of 4 percent for Tier-1 capital. The large increase in market-value-based capital ratios during 1994–2000 is probably a reflection of the high levels of the overall stock market during that time.

In *Figure 10.6*, we show the number of bank failures during 1992–2004. As is evident, the number of failures decreased dramatically during the early 1990s and has stayed relatively low as banks have operated with healthy capital ratios.

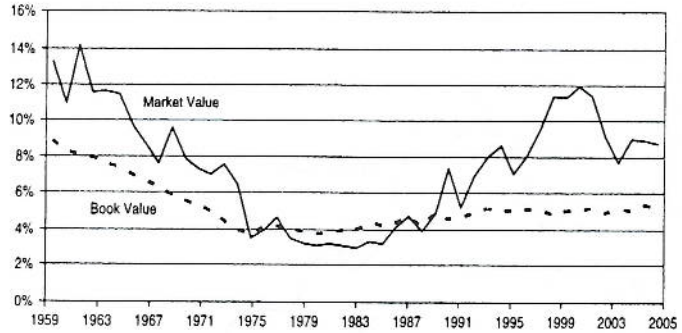


FIGURE 10.4 Capitalization of the 25 Largest Bank Holding Companies During 1959–2005
This Figure shows the weighted average capitalization of the 25 largest bank holding companies from 1959–2005. Capital ratios are expressed in book values (shareholder equity/total assets) and market values (market value of equity/market value of assets).
Source: Compustat and own calculations.

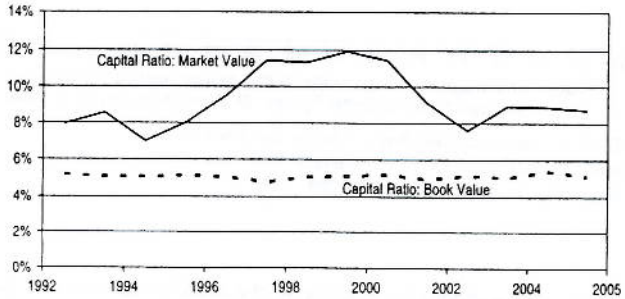


FIGURE 10.5 Capitalization of the 25 Largest Bank Holding Companies During 1992–2005
Source: Compustat and own calculations.

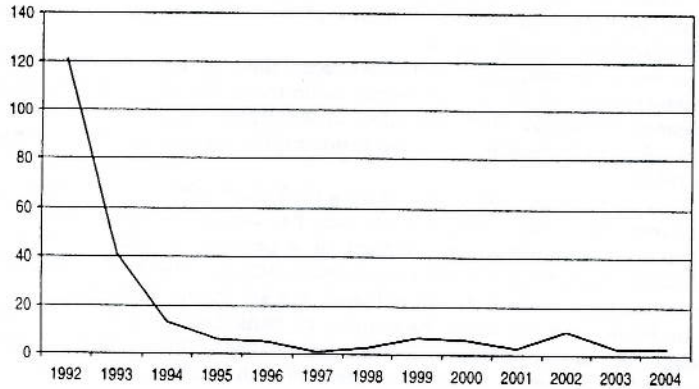


FIGURE 10.6 Bank Failures 1992–2004
Source: FDIC Historical Statistics on Banking: Closings and Assistance Transactions.

The Great Deposit Insurance Debacle

General Background

We have now reviewed both the theory and some empirical evidence about the effects of deposit insurance on depository institutions' risk-taking behavior. In trying to understand the great deposit insurance debacle of the 1980s, it is important to remember that until the mid-1970s deposit insurance worked remarkably well. But, two developments undermined federal deposit insurance. One is the lowering of bank charter values, which increased managers' incentives to take more asset risk and to also engage in fraud. The other is the decline in regulatory vigilance over the same period; this simply exacerbated the moral hazard problem of federally insured depository institutions.

The waste that resulted from the collapse of the thrift industry and the many banking failures in the 1980s can be classified into three categories: excessive risk-taking, excessive consumption of perquisites by top executives, and outright fraud. Moreover, these diversions/destructions of wealth were possible due to three factors working in concert: deposit insurance with risk-insensitive pricing, low charter values due to deregulation, and lax monitoring by regulators. This laxity in monitoring, caused by a lowered commitment of resources to supervision, was also compounded by cozy relationships between some regulators and the institutions they were supposed to be watching over. In *Figure 10.7*, we have provided a simple schematic to summarize these effects.

It is not as if S&L and bank managers woke up one morning in the 1980s and decided to change the way they made decisions in order to "rip off" the taxpayers.

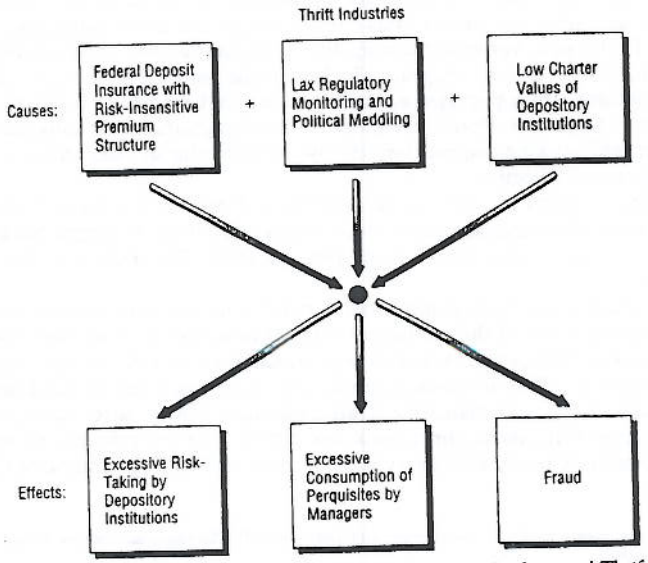


FIGURE 10.7 Schematic of Effects Responsible for Problems in Banking and Thrift Industries

The point is that their *incentives* were altered. Their *decision rule* was still the same, but the altered incentives changed their behavior. The reasons for the deposit insurance crisis can therefore be traced not just to the managers of depository institutions, but also to the politicians and regulators who pursued myopic and hasty policies. In what follows, we briefly discuss the causes and effects depicted in *Figure 10.7*.

Regulatory and Political Culpability

For some years, S&L regulators tried to ignore problems in the thrift industry. Hoping that problems would improve, regulators permitted insolvent institutions to continue to operate. Our analysis of credit risk in Chapters 5 and 6 highlighted the important incentive effects of capital on a borrower's risk-taking propensity. The same is true of depository institutions; their propensity to take risk is greater when capital is lower. When capital is negative, excessive risk-taking is easy to predict. Indeed, the Federal Home Loan Bank Board (FHLBB) was quite aware that the thrift industry was in deep trouble in 1981, but chose not to close all the insolvent institutions.[28]

This inaction was part of a broader regulatory malaise. The main findings of a 2-month *USA Today*-Gannett News Service investigation are listed below.[29]

- Some regulators had close ties to the industry.[30]
- In some cases, *regulators* suggested that S&Ls try to grow rapidly and to invest in risky ventures as a way of quickly boosting profits.
- Regulatory agencies lacked the powers and the human resources to monitor rapidly growing S&Ls.
- Congress repeatedly refused requests to add S&L examiners, and told the FHLBB to go easy on problem S&Ls.[31]

Excessive Risk-Taking

As previously discussed, the three prominent ways to detect excessive risk-taking involve examining capital-to-total-asset ratios, interest rates on large (uninsured) CDs, and the assets in the institution's portfolio. The last two are briefly discussed below.

28. A 1990 issue of *The American Banker* quoted Mr. Richard T. Pratt, then chairman of the FHLBB, "Had we liquidated the S&L industry in 1981, it would have cost $178 billion–$380 billion in today's dollars. It would have been the most foolish public policy that could have possibly been undertaken."

29. See *USA TODAY*, February 14, 1989.

30. Mr. Tom Huston, former Iowa state banking superintendent, claims that regulators traveled too much at industry expense. He said, "They were so loved and so well-treated ... that no wonder they couldn't make a rational decision."

31. Mr. Edwin Gray, FHLBB chairman from May 1983 to June 1987, blames Congress and the Reagan administration for failing to give regulators more power, and he blames the powerful S&L lobby for influencing them. In *USA TODAY* (2/14/1989), Mr. Gray was quoted as saying the following: "We were asking Congress and the Reagan administration for help and getting nothing. We had a rag-tag bank of 700 examiners, who were expected to monitor $1 trillion in assets and 3,300 S&Ls. Sometimes our examiners were hired away by the S&Ls they were examining." It turns out that entry-level examiners were paid $14,000 per year during this time, and the turnover rate was 25 percent.

Higher Interest Rates on CDs: Riskier institutions must pay higher interest rates on large CDs, or conversely, those institutions that offer to pay higher interest rates on their deposits anticipate investing in high-risk, high-yield assets to cover their deposit funding costs. Since risk-taking incentives are the strongest in insolvent and nearly insolvent institutions, one would expect such institutions to be paying the highest rates. This is precisely what happened in the southwestern United States, where the S&L industry was devastated. Higher interest rates offered by insolvent institutions led to a self-fulfilling prophecy. When a depository institution has low net worth, it is expected to invest in riskier assets, so that depositors demand relatively high interest rates. These high interest rates, in turn, increase the attractiveness of high-yield, high-risk assets to the institution, thus completing the cycle. Moreover, in order to compete with insolvent institutions, solvent institutions may be compelled to offer higher interest rates on their deposits, leading to stronger incentives to invest in riskier assets.

Investments in High-Risk Assets: There is ample evidence of excessively risky investments by S&Ls. These investments included loans to developers to build ski resorts, speculative positions in government securities, junk bond portfolios, and so on.

Excessive Consumption of Perquisites by Managers

Although it is empirically difficult to determine whether a given level of perquisites consumption by a manager is "appropriate," some of the examples are striking and suggestive of abuse. These include institutional purchases of planes to transport top managers from their places of residence to their offices, payments for escort services, offices lined with expensive antiques and paintings, and gold-painted toilets. Many of the institutions where such apparent abuse occurred were investigated by the FSLIC.

Fraud

Estimates of *direct* losses to the government due to fraud by S&L managers range from $8 billion to $15 billion, and fraud is suspected in 80 percent of failed S&Ls. Parties, mansions, airplanes, women, Rolls-Royces, and Cayman Island bank accounts are some of the perks that S&L executives showered upon themselves as they looted federally insured deposits.

The S&L crooks also caused failures of S&Ls run by honest managers, by selling them stakes in their bad loans. For example, the now-insolvent First Federal Savings and Loan of Malvern, Arkansas, bought an interest in a doomed $44 million loan to a high-rise condo in Honolulu, which subsequently defaulted.

Many of the fraud cases are very complex. Shady S&Ls and equally shady borrowers combined dozens of loans, companies, and properties into convoluted deals to cover personal use of S&L deposits. Some S&Ls made borrowers pay big one-time fees—4 percent to 10 percent of the loan—in order to obtain loans. The S&Ls would report these fees as income, which boosted profits. Many loans were never repaid, leaving the property in the S&L's hands. An S&L executive might get a kickback for participating in the scheme. In Texas, this strategy was described as: "Heads, I win. Tails, FSLIC loses."

Following the S&L debacle, the government has filed approximately 100,000 civil suits against S&L executives, directors, owners, borrowers, and others believed responsible for contributing to the insolvency of S&Ls. The success of these prosecution efforts, and of attempts to recover some of the losses due to fraud, negligence, and simple mismanagement, remains uncertain.

To summarize, the greatest banking debacle since the Great Depression was *not* just an "unfortunate break" or an outcome of exogenous changes in the banking environment. Increasing competition increased interest-rate volatility and deregulation reduced the profitability of depository institutions, substantially diminishing charter values. Models of bank behavior predict increased risk-taking by federally insured institutions in such a setting, suggesting a need for improved regulatory monitoring. Unfortunately, safety was sacrificed at the same time that the industry was deregulated, as resources devoted to regulatory supervision were decreased. Regulatory ineptness and political meddling compounded the effects of poorly-thought-out initiatives.[32]

Banking Fragility, Deposit Insurance and Developments Since the Great Deposit Insurance Debacle

We have seen in this chapter that deposit insurance induces moral hazard and invites banks to engage in reckless risk-taking. That is, there is an inherent paradox in the use of deposit insurance as a way to diminish the likelihood of bank runs and banking fragility. The safer banks feel due to deposit insurance, the greater is their risk-taking propensity! It is for this reason that it may be socially efficient to impose a limit on the level of deposit insurance, thereby leaving room for market discipline, which then opens up the possibility of bank runs and banking fragility. In other words, there may be an "optimal" amount of banking fragility that strikes the right balance between the market discipline associated with the possibility of bank runs to temper banks' risk-taking incentives and the need to ensure that the likelihood of runs is not so high as to make banking excessively fragile.[33]

One could argue that one way to cope with deposit insurance–related moral hazard is to use capital requirements as an instrument to reduce banks' proclivity to take excessive risk. Regulatory reforms associated with the Basel I and Basel II capital accords and FDICIA of 1991 (see the next two chapters) lend strong support to the hypothesis that sufficiently high capital requirements can be effective in controlling risk.[34] The incidence of bank failures during the 1990s and 2000–2005 has been remarkably low, and the FDIC has been building up its reserves.

The real question, however, is whether we need deposit insurance in the first place, for a lot of the regulatory apparatus we observe would be unnecessary were it

32. We recommend reading Adams (1990), Mayer (1990), and White (1991) for accounts of the many factors that contributed to the implosion of the thrift industry in the United States.

33. This implication can be drawn from Calomiris and Kahn (1991). It is a point that has been made by Diamond and Rajan (2001). For other analyses of banking fragility, see Allen and Gale (2001).

34. Coval and Thakor (2005) explain that a certain level of bank capital may even serve the purpose of ensuring the viability of a bank that seeks to serve as a "bridge" between borrowers and savers with divergent beliefs.

not for deposit insurance.[35] But would we not have excessive bank runs without deposit insurance? This is actually an open question. Mutual funds have no deposit insurance and we have not observed any runs. At the end of the day, a fundamentally sound banking system, backed up by a credible lender of last resort, may not be as fragile without deposit insurance today as it may have been in the past. If this is true, the entire system of deposit insurance and regulations may have to be reconsidered.

Conclusion

We have devoted this chapter to an extensive discussion of the deposit contract, liability management, and deposit insurance. The nature of the deposit insurance contract is such that it leaves the bank vulnerable to runs, and the banking system vulnerable to panics. It appears that deposit insurance served its purpose of minimizing bank runs and panics. Indeed, for almost 50 years since the inception of federal deposit insurance in 1933, failure rates in the banking and thrift industries have been abnormally low compared to other industries. Moreover, this stable environment meant that liability management was not a pressing issue for banks.

But all that changed in the 1970s and 1980s. As interest-rate volatility increased and interest-rate restrictions were relaxed and then eliminated, liability management became a significant concern for banks. Moreover, a combination of deregulation, heightened volatility in market prices, lax regulatory monitoring, political interference, and corrupt executives in federally insured institutions significantly undermined the safety of the industry, and imposed monstrous losses on the deposit insurance funds. It is somewhat ironic that these events were quite predictable, in light of what was known *prior* to these events. Regulatory reforms that followed have helped to significantly improve banking stability.

Review Questions

1. What are the main economic features of the demand deposit contract and how do these features discipline management when deposits are uninsured?
2. What measures were used to cope with bank runs and panics prior to federal deposit insurance? Why were these not entirely satisfactory?
3. What is a bank run and how can you explain a run on *economic* grounds?
4. How does deposit insurance prevent runs and panics?
5. Explain the similarity between deposit insurance and a common stock put option and how this leads to moral hazard.
6. Why did deposit insurance work so well in the United States until 1980 despite the obvious moral hazard, and why did it fail after that?
7. Discuss the roles of bank managers, accountants, regulators, and politicians in the "great banking/S&L debacle."
8. What is liability management and what are its main objectives?
9. What is the agency problem between the shareholders and managers of a bank in liability management?

35. See Miller (1995) for a forceful argument in favor of dismantling federal deposit insurance.

10. Is moral hazard unethical, illegal, or neither? Can you outline a conceptual framework for defining unethical behavior by a depository institution?
11. What aspects of S&L/bank behavior would you consider unethical? For example, were junk bond investments unethical? Be sure to take an *ex ante* perspective and not use "20-20 hindsight."
12. Why do you think unethical behavior became so rampant in the last decade and not prior to that? Did people change? Did morals decline in general? Did the environment change? Can you relate unethical behavior during this period to similar behavior during other periods in history?
13. Consider a bank that receives a $1 deposit from each of 200 different depositors at $t = 0$. It invests $25 of shareholders' equity in the bank and lends $200, keeping $25 as cash reserves. Out of the 200 depositors, there are 75 depositors (called type-D_1 depositors) who are capable of monitoring the bank's management; the remaining depositors (called type-D_2 depositors) have kept their money in the bank simply for transactions and safekeeping. The cost of monitoring the bank for an individual type-D_1 depositor is $0.03 per period.

    The bank has two mutually exclusive investment opportunities. Project (or loan) A pays $300 with probability 0.6 and zero with probability 0.4 at $t = 1$. Project B pays $250 with probability 0.8 and $220 with probability 0.2 at $t = 1$. If the bank chooses one of these two projects, the probability that the bank will actually end up with that project is 0.7. With probability 0.3, the bank will have inadvertently chosen the other project. Thus, we assume that the bank may make errors in project choice. By monitoring the bank, a type-D_1 depositor can discover the bank's true project choice at some point in time intermediate between $t = 0$ and $t = 1$, say at $t = 1/2$. These depositors can, if they desire, force liquidation of the bank by withdrawing their deposits at $t = 1/2$. Note that the bank's loans/projects mature at $t = 1$. If they are liquidated at $t = 1/2$, they are worth only $70 to the bank. Under the terms of the deposit contract, the bank promises to pay 15 percent interest (conditional on the bank having the financial capacity to do so) if deposit withdrawal occurs at $t = 1$, and no interest if withdrawal occurs before that. Thus, a depositor is entitled to $1.15 if she withdraws at $t = 1$, and $1 if she withdraws at $t = 1/2$. The risk-free discount rate is zero and all agents are risk neutral.

    All the type-D_2 depositors plan to withdraw at $t = 1$, but each is subject to a random liquidity-motivated desire to withdraw at $t = 1/2$. To simplify, we will assume that even though no one knows in advance which (type-D_2) depositors will wish to withdraw at $t = 1/2$, the fraction of those who will wish to withdraw is known to be 25/125. That is, 25 type-D_2 depositors will wish to withdraw at $t = 1/2$. Assume that the bank's managers make decisions in the best interests of their shareholders. Compute the equilibrium strategies for the bank and its depositors.

References

Adams, James Ring, *The Big Fix: Inside the S&L Scandal*, John Wiley & Sons, Inc., New York, 1990.

Allen, Franklin, and Douglas Gale, *Comparing Financial Systems*, Cambridge, MA: MIT Press, 2001.

Association of Reserve City Bankers, *Capital Issues in Banking*, Chicago, December 1988.

Black, Fisher, and Myron Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* 81, 1973, 637–659.

Bryant, John, "A Model of Reserves, Bank Runs, and Deposit Insurance," *Journal of Banking and Finance* 4, 1980, 335–344.

Calomiris, Charles W., and Charles M. Kahn, "The Role of Demandable Debt in Structuring Optimal Banking Arrangements," *American Economic Review* 81–3, June 1991, 497–513.

Calomiris, Charles W., Charles M. Kahn, and Stefan Krasa, "Optimal Contingent Bank Liquidation Under Moral Hazard," unpublished manuscript, April 1991.

Chan, Yuk-Shee, Stuart I. Greenbaum, and Anjan V. Thakor, "Is Fairly Priced Deposit Insurance Possible?" *Journal of Finance* 47–1, March 1992, 227–246.

Chari, Varadarajan V., and Ravi Jagannathan, "Banking Panics, Information, and Rational Expectations Equilibrium," *Journal of Finance* 43, July 1988, 749–761.

Clair, Robert T., "Deposit Insurance, Moral Hazard, and Credit Unions," *Economic Review*, Federal Reserve Bank of Dallas, July 1984, 1–8.

Coval, Josh, and Anjan V. Thakor, "Financial Intermediation as a Beliefs-Bridge Between Optimists and Pessimists," *Journal of Financial Economics* 75–3, March 2005, 535–570.

Diamond, Douglas, and Raghuram Rajan, "Liquidity Risk, Liquidity Creation and Financial Fragility: A Theory of Banking," *Journal of Political Economy*, April 2001.

Diamond, Douglas W., and Philip Dybvig, "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy* 91, June 1983, 401–419.

Flannery, Mark, "Debt Maturity and the Deadweight Cost of Leverage: Optimally Financing Banking Firms," working paper, University of Florida, 1992.

——, "Pricing Deposit Insurance When the Insurer Measures Bank Risk with Error," *Journal of Banking and Finance* 15, 1991, 975–998.

Gibbons, J.S., *The Banks of New York, Their Dealers, the Clearing House, and the Panic of 1857*, Greenwood Press (reprint of 1859 original), New York, 1968.

Gorton, Gary, "Banking Panics and Business Cycles," *Oxford Economic Papers* 40, 1988, 751–781.

Gorton, Gary, and Donald J. Mullineaux, "The Joint Production of Confidence: Endogenous Regulation and Nineteenth Century Commercial-Bank Clearinghouses," *Journal of Money, Credit and Banking* 19–4, November 1987, 457–468.

Hutchison, David E., and George G. Pennacchi, "A Framework for Estimating the Value and Interest Rate Risk of Retail Bank Deposits," working paper 92–30, Federal Reserve Bank of Chicago, December 1992.

Jacklin, Charles, "Banks and Risk-Sharing: Risk-Sharing: Instabilities and Coordination," in S. Bhattacharya and G.M. Constantinides (eds.), *Financial Markets and Incomplete Information*, Rowman and Littlefield, Totowa, N.J., 1989.

——, "Demand Deposits, Trading Restrictions, and Rísk Sharing," in E.C. Prescott and N. Wallace (eds.), *Contractual Arrangements for Intertemporal Trade*, University of Minnesota Press, 1987.

Jacklin, Charles, and Sudipto Bhattacharya, "Distinguishing Panics and Information-Based Bank Runs: Welfare and Policy Implications," *Journal of Political Economy 96*, June 1988, 568–592.

Kahane, Yehuda, "Capital Adequacy and the Regulation of Financial Intermediaries," *Journal of Banking and Finance* 1, June 1977, 207–218.

Keeley, Michael C., "Deposit Insurance, Risk, and Market Power in Banking," *American Economic Review* 80, December 1990, 1183–1200.

The Theory of Corporate Finance

Jean Tirole

Princeton University Press
Princeton and Oxford

Published by Princeton University Press,
41 William Street, Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press,
3 Market Place, Woodstock, Oxfordshire OX20 1SY

Library of Congress Cataloguing-in-Publication Data

Tirole, Jean.
The theory of corporate finance / Jean Tirole.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-691-12556-2 (cloth: alk. paper)
ISBN-10: 0-691-12556-2 (cloth: alk. paper)
1. Corporations—Finance. 2. Business enterprises—Finance.
3. Corporate governance. I. Title.

HG4011.T57 2006
338.4'3'001—dc22 2005052166

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

This book has been composed in LucidaBright and
typeset by T&T Productions Ltd, London

Printed on acid-free paper ∞
www.pup.princeton.edu

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10

2

Corporate Financing: Some Stylized Facts

2.1 Introduction

One of the goals of corporate finance theory is to help predict or advise on security issues and payout policies at various stages of a firm's life cycle. There is much discretion involved in specifying a security's cash-flow rights, control rights, and other rights (collateral, options) and the contingencies under which these rights are triggered and exercised. As for corporate governance in Chapter 1, the purpose of this selective review of corporate financing and payout policies is to guide the later theoretical construct and to enable future feedback concerning the accuracy of its predictions.

This chapter offers a succinct description of the financing of firms, focusing on their main financial instruments: debt and equity, in their different varieties.

2.1.1 A Wide Variety of Claims

The simplest form of debt is a claim to a predetermined level on the firm's income. Equityholders receive any profit, that is, are "residual claimants," beyond that level. On the other hand, if debt is not repaid, shareholders receive nothing and debtholders are entitled to the existing income. The view of debt and equity as claims with concave and convex return structures, respectively, is represented in Figure 2.1 for some arbitrary reimbursement level D.

Note that debt in a highly leveraged or "undercapitalized" firm (D high) resembles equity in a modestly leveraged or "well-capitalized" one (D low), in that in both cases claimholders are basically residual claimants at all income levels. Thus, securities that are labeled one way (e.g., debt) may have cash-flow features (and, as we will later see, functions) that are more characteristic of another type of securities (e.g., equity).

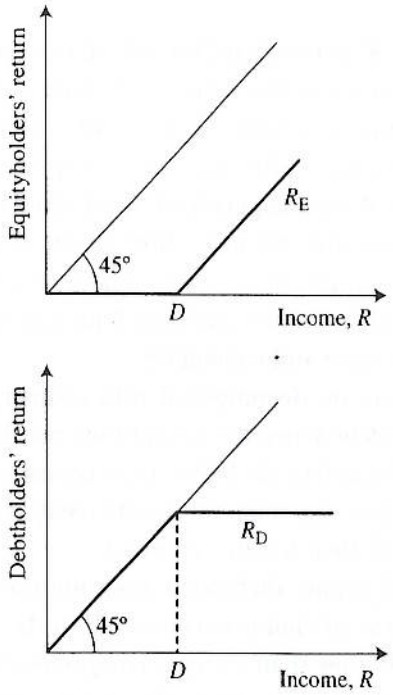


Figure 2.1

This elementary description of financial claims is a useful starting point, but it is oversimplistic. In particular, it ignores the following considerations:

- The firm is usually an ongoing entity, which produces a stream of returns rather than a single one. The one-dimensional representation of Figure 2.1 is at best a condensed view of the stream of returns attached to the claim.
- Who holds the claim in general matters. Corporate governance, for example, depends on whether equity is held by "insiders" (managers, entrepreneurs) or by "outsiders"; on whether share ownership among outsiders is concentrated in the hands of one or a couple of main shareholders or is spread among many shareholders; and on whether debt is held by a large player (such as a bank) or by dispersed investors.

• Claims are not simply defined by their attached returns streams. Claimholders also receive control rights, that is, the right to make decisions, whose scope is either specified in advance or is defined by default (residual rights of control), in circumstances that are defined contractually. For example, shareholders usually have control rights as long as debt covenants are satisfied, but debtholders acquire some control rights in case of violation of these covenants.

• Income (R) may be hard for outsiders to verify in the case of small entrepreneurs. Medium and large firms in contrast usually have a fairly reliable accounting structure, although accounting manipulations may enable managers to shift reported income between years (for instance, through the choice of date of recognition of expenses and revenue), and more generally to distort the overall picture of earnings performance and capabilities.

• Debt may be decomposed into ordinary debt and *secured debt*. When debt is not fully reimbursed, secured debtholders do better than ordinary debtholders as they can seize the assets used as collateral as part of their lending contract.

• The debt-equity dichotomy does not do justice to the richness of claims encountered in the corporate world. Rather than giving a comprehensive description of the many existing claims,[1] here we shall describe a few of the most common intermediate claims between debt and equity.

First, one must distinguish between *senior debt* and *subordinated* or *junior debt*. In the case of default, more senior debtholders are reimbursed first; holders of subordinated debt are then repaid if enough is left, as they have priority over equityholders. Junior debt must therefore deliver a higher yield than senior debt in order to compensate for the higher risk of default. Figure 2.2 depicts the returns attached to subordinated debt when the firm must pay D to senior debtholders and d to junior debtholders. The return schedule for subordinated debt is neither convex nor concave. For d large, subordinated debt resembles equity: a severely undercapitalized (that is, highly leveraged) firm is unlikely

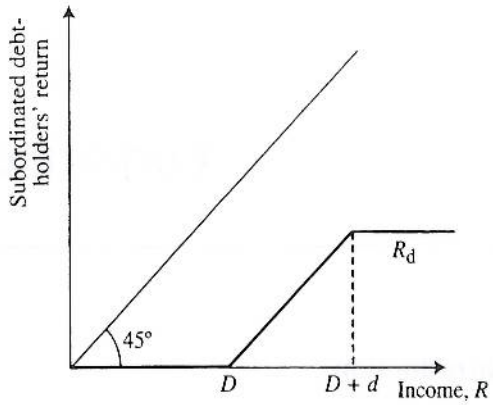


Figure 2.2

to produce much income for its shareholders, so the holders of subordinated debt are almost residual claimants once senior debt is reimbursed. Conversely, for small amounts of senior debt D, the preferences of junior debtholders resemble those of ordinary debtholders.

Another common intermediate claim is *(cumulative) preferred stock*. Preferred stock is like debt in that its holders are entitled to a fixed, predetermined repayment. Unlike debt, the firm is not obliged to pay back this specified amount, and thus nonrepayment does not trigger default. However, the firm cannot pay a dividend on (common) stock unless the cumulative (past and current) payments due to preferred stockholders have been made. Preferred stockholders are thus senior to (common) stockholders. Also, while common stocks usually carry voting rights, preferred stockholders often do not have voting rights. They thus have little control over the firm. Their claim is junior to debt, and so for a financial structure made of debt, preferred stock, and equity the returns attached to preferred stocks are also depicted by Figure 2.2 in a single-period context. However, in an ongoing context, preferred stock gives the firm more flexibility on the repayment schedule than subordinated debt.

Subordinated debt and preferred stocks are instances of *mezzanine finance*, that is, of investments that occupy a middle-level position between common equity and senior debt in the firm's capital structure. Mezzanine investments[2] (with exceptions: preferred stocks are usually publicly traded)

1. See, for example, Allen et al. (2005) for more details. Finnerty (1993) provides an overview of some sixty recently introduced types of (debt and equity) security.

2. See Willis and Clark (1993) for more on mezzanine finance.

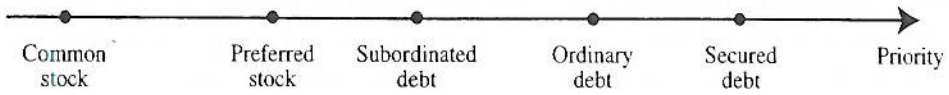


Figure 2.3 Priority structure.

generally are privately placed[3] and often include equity participations in the form of warrants[4] and stock appreciation rights.[5]

The priority structure of the main claims described so far is summarized in Figure 2.3.

A last major intermediate claim is *convertible debt*, one of the many claims that take the form of an option, which the holders can elect to exercise if circumstances are favorable. Convertible debt is basically debt, except that its holders can exchange it for the firm's shares at some predetermined conversion rate.[6] The holders of convertible debt may exercise this option and acquire shares, for instance, if the firm's prospects become favorable, or if for a given expected income of the firm the riskiness of the firm's income has increased due to changes in the environment or to managerial choices (well-diversified holders of a convex, respectively concave, claim like, respectively dislike, risk). Indeed, Jensen and Meckling (1976), among others, have argued that the convertibility option protects debtholders against excessive risk taking by the firm. To see why, consider a corporate move that does not affect the firm's expected profit, but increases its riskiness.[7] For example, the firm may put all its eggs in the same basket by investing in a single risky activity, or by refraining from hedging against market risk (e.g., foreign exchange, interest rate, or raw material risk). Risk-neutral or well-diversified investors benefit from this increase in risk if they hold a convex claim, and they lose if their claim's return profile is concave. In this sense, (diversified) equityholders like (mean-preserving) increases in risk while debtholders dislike such increases in risk. Indeed, equityholders may gain even if the increase in riskiness reduces total investor value (value of debt plus equity), the case of a mean-decreasing increase in risk. For this reason, debtholders are particularly wary of decisions that affect riskiness. To protect themselves against abusive risk taking by the corporation, debtholders may demand covenants that force the firm to exert care; but it may be difficult to force the firm to hedge adequately and so the debtholders may be further protected by a convertibility option: a move that enriches shareholders to the detriment of debtholders is then undone if the latter have the option to convert their claim into an equity claim.

2.2 Modigliani-Miller and the Financial Structure Puzzle

Why do we care about the firms' financial structure? The short answer is that insiders as well as outsiders (commercial banks, investment banks, rating agencies, venture capitalists, equityholders, etc.) devote a lot of attention to its design. But we must also ask whether this attention is warranted. As a matter of fact, economists were stunned when, in two articles in 1958 and 1961, Modigliani and Miller came up with the following rather striking and somewhat counterintuitive result. Under some conditions, the total value of the firm—that is, the value of all claims over the firm's income—is independent of the financial structure. That is, the level of debt, the split of debt into claims with different levels of collateral and different seniorities in the case of bankruptcy, dividend distributions, and many other characteristics or policies relative to the financial structure have no impact on total value. In other words, decisions concerning the financial structure affect only how

3. A private placement is an issue that is offered to a single or to a few investors. In the United States, private placements do not have to be registered with the SEC.

4. A warrant is a long-term call option, that is, an option to buy the security at a specific exercise price on or before a specified exercise date.

5. Stock appreciation rights are stock options which enable their holder to receive the capital gain relative to the exercise price without supplying cash.

6. A convertible bond resembles a package of a bond and a warrant (a warrant is an option to buy shares at a set price on or before a given date). The difference is that the payment to buy the shares is in cash in the case of a warrant, and in a bond in the case of a convertible.

7. In the sense of a mean-preserving spread (i.e., second-order stochastic dominance).

the "corporate pie" (the statistical distribution of income that the firm generates) is shared, but has no effect on the total size of the pie. Thus, an increase in debt or a dividend distribution dilutes the debtholders' claim and benefits the shareholders, but the latter's gain exactly offsets the former's loss.

To illustrate this point, consider the simple debt-equity structure of Figure 2.1, and assume that investors are risk neutral.[8] Let V_E and V_D denote the values of equity and debt for debt repayment D. Then the total value,

$$\begin{aligned} V_E + V_D &= \mathcal{E}(\max(0, R - D)) + \mathcal{E}(\min(R, D)) \\ &= \mathcal{E}(R), \end{aligned}$$

is independent of D, where $\mathcal{E}(\cdot)$ denotes the expectation with respect to the distribution of the random variable R.[9]

Add to this result the observation that efficient corporate policies should aim at maximizing the size of the corporate pie: any increase in the firm's total value brought about by a change in policy can be divided among the claimholders in a way that makes everyone better off.[10] Modigliani and Miller's conclusion then follows: the financial structure is irrelevant. Managers and investors might as well devote their time to more useful tasks and simplify their financial structure by issuing a single claim, which could be labeled "100% equity" or "equity without debt" (this is the claim depicted by the 45° line in Figure 2.1). The firm would then become an "all-equity firm."

Similarly, the payout policy (dividends and share repurchases/issuance) has no impact on firm value. To illustrate this, consider an all-equity firm, again with risk-neutral investors. Time is discrete: $t = 0, 1, 2, \ldots$. In each period t, a random net revenue R_t accrues; then a per-share dividend d_t is paid, the number of shares is adjusted from n_{t-1} to n_t, and an investment I_t is sunk.[11] Consider, for each t, a given (state-contingent) investment policy I_t, as well as an (also state-contingent) choice of dividend d_t and number of shares n_t ($n_t < n_{t-1}$ in the case of share repurchases, $n_t > n_{t-1}$ when new shares are issued). Let P_t denote the price of a share at the end of period t (after the dividend payment) and β the discount factor.

By arbitrage,

$$P_t = \beta \mathcal{E}[d_{t+1} + P_{t+1}].$$

Furthermore, at date t, there is an accounting equality between the sum of revenue and amount raised in the capital market (this amount is negative for share repurchases) and the sum of dividend and investment:

$$R_t + P_t(n_t - n_{t-1}) = n_{t-1} d_t + I_t.$$

The total value of shares in the firm at the end of period t is therefore

$$\begin{aligned} V_t \equiv n_t P_t &= \beta n_t \mathcal{E}[d_{t+1} + P_{t+1}] \\ &= \beta \mathcal{E}[R_{t+1} - I_{t+1} + (n_{t+1} - n_t) P_{t+1} + n_t P_{t+1}] \\ &= \beta \mathcal{E}[R_{t+1} - I_{t+1} + V_{t+1}] \\ &= \mathcal{E}\left[\sum_{\tau \geq 1} \beta^\tau (R_{t+\tau} - I_{t+\tau}) \right] \end{aligned}$$

by induction. Thus, the value of claims on the firm depends only on its "real" characteristics—investment policy and net income—and not on the dividend and capital market choices.

It is only recently that economists have started developing a better understanding of the role of the financial structure. And, although the theory of corporate finance is still evolving, it is fair to say that considerable progress has been made. To examine whether the business community's close attention to the financial structure is warranted, economists have questioned the idea that the size of the pie is exogenously determined. At an abstract level, one can analyze the matter in the following terms. Whenever managerial decisions cannot be perfectly specified contractually, the incentives given to those who pick those decisions affect the firm's income (the

8. The Modigliani-Miller irrelevance result is much more general than this. In particular, it holds even if investors are risk averse (the proof then employs "state-contingent prices").

9. Risk neutrality is not required for the result. Intuitively, with risk-averse investors, one can still define "state-contingent prices," that is, the prices of 1 unit of income in the various states of nature, and apply this equality to the sum of the values of equity and debt.

Also, the notation for expectations will be $E[\cdot]$ in the rest of the book. We use another notation here in order to avoid a confusion with equity.

10. Unless the winners do not have enough money, or more generally means of exchange, to compensate the losers (on this, see Chapter 3).

11. The investment, together with previous investments, will generate a random income R_{t+1} through a production function that we do not need to describe here.

size of the pie) and therefore the split of the pie matters. To clarify this point, consider the numerous decisions taken by the firm's "insiders," namely, the entrepreneurial or managerial team. As discussed in Chapter 1, there is no *a priori* reason why insiders have proper incentives to maximize total firm value. Casual observation suggests that managers do not always exert enough care in their choice of projects or in their supervision of divisions and subsidiaries; that they may waste corporate funds to build empires; that they sometimes select policies because they are easy to implement or will not jeopardize a comfortable managerial position; that some divest resources to indulge in perks (luxurious headquarters, entertainment expenses, corporate jets); or that they may select suppliers or employees on grounds (e.g., friendship) other than efficiency.

Such hazards have been known for a long time, and "governance structures" have been put in place that limit (but do not eliminate) deviations from profit maximization. As discussed in Chapter 1, there are roughly three ways of preventing insiders' misbehavior. First, some contractual constraints can be imposed on managers in the form of covenants and other clauses in financial deals. However, covenants by nature can be based only on public and therefore coarse information, and have their limits. Second, claimholders and managers can agree to build strong or "high-powered" managerial incentives to maximize profit. As pointed out in Chapter 1, though, the provision of high-powered incentives to entrepreneurial or managerial teams is costly, and is unlikely per se to achieve perfect congruence between insiders' and outsiders' interests. It is important that such incentives, if any, be complemented by monitoring and occasional intervention by outsiders: deviations from profit maximization may be detected by outsiders, who can put the firm back on track if they have the authority to do so. Because monitoring is partly a public good for claimholders and therefore is likely to generate free riding, a ubiquitous pattern in efficient corporate financing is the implicit or explicit delegation of monitoring to one or several claimholders with large enough stakes in the firm to induce them to monitor managerial policies, and with a contractual right to interfere if management goes awry. The monitoring patterns differ in their intensity and in the nature of the monitors' claims. Again from Chapter 1, we know that monitors may have debt claims (commercial banks and insurance companies, investment banks), equity claims (large shareholder, such as a pension fund, another corporation, a venture capital firm, or an LBO specialist), or no claim at all (rating agencies, whose incentives are purely based on their reputation to grade corporations accurately).

Our presentation of the main stylized facts about corporate financing emphasizes informational and control issues, which we feel are central to a good understanding of the matter. This does not mean that other considerations, such as tax or clientele effects, are irrelevant. *Tax considerations* influence the choice of financial structure. In particular, debt usually enjoys tax advantages relative to equity; relatedly, junk bonds, which are highly risky bonds, may be issued partly to avoid the corporate income tax that is borne by equity. Taking advantage of the imperfections of the tax system is a consummate and perennial exercise for financial experts (as well as for other experts), but its details are often country- and time-specific, so we will ignore them here.[12] Another important consideration is the presence of *clientele effects* in the supply side of loans. Many financial intermediaries (banks, insurance companies, pension funds, mutual funds) are subject to regulatory requirements, which penalize them for holding certain types of asset or even prohibit them from doing so.[13] The motivation for such controls is that financial intermediaries are subject to moral hazard just like nonfinancial companies, the effect of which is explored further in Chapter 13. Issuers of claims respond, of course, to the fact that financial intermediaries (the main purchasers of the claims) have for regulatory reasons higher demands for certain classes of claims.

A third consideration relates to the *enforcement of financial contracts.* We will mostly assume that such contracts are enforced. In practice, bankruptcy law may not always respect agreements and may

12. See the introduction to the book for a few references on the impact of taxes on financial structures.

13. For more institutional details as well as for a comparison between the governance structures of nonfinancial and financial companies, see Chapters 2 and 3 in Dewatripont and Tirole (1994a).

reshuffle the claims. For example, some bankruptcy laws are prejudiced against secured debt and do not fully allocate the collateral to secured debtholders. Bankruptcy laws can therefore have an impact on the financial structure of firms.[14]

The chapter is organized as follows. Section 2.3 considers debt claims and classifies them along several dividing lines: public versus private, secured versus unsecured, high- versus low-intensity monitoring, priority, covenants. Section 2.4 performs a similar analysis with respect to equity claims. Section 2.5 looks at the firm's actual financial choices, and asks the following questions: How are new investments financed? What are the determinants of leverage? Which firms are financially constrained? How are financial structures affected by business cycle-related fluctuations and by the firm's profit realizations?

2.3 Debt Instruments

A prospective borrower faces a number of choices. First, the firm must choose from whom to borrow. It can apply for a bank loan, place debt privately with institutions such as life insurance companies, issue bonds to the public at large, or use still other forms of credit such as trade credit (that is, credit from suppliers). Second, the firm can issue short-term (possibly rolled over) debt or long-term debt. Third, it can restrict its flexibility in future decision making and transfer some control rights to lenders through the writing of covenants. Fourth, it can pledge assets as collateral. And, fifth, the firm can establish a structure of priority among debt instruments in case of default.

A typical debt liability specifies:[15]

- the amount of borrowing (the principal), the term (maturity), the rate of interest, the scheduling (whether the amount borrowed is due only at maturity or a specified portion of the issue is retired each year—the case of a "sinking fund" requirement), and possibly other conditions (indexation, call provision,[16] etc.);
- a mechanism for transmitting timely, credible information to the lender(s);
- warranties (in which the borrower confirms in writing the accuracy of information about the legal status of the firm, its financial statements, the absence of pending or threatened litigation against it, the absence of previous lien on the collateral or of unpaid taxes, etc.);
- affirmative covenants, which force the borrower to take actions that protect the lender(s);
- negative covenants, which place restrictions on the borrower's ability to take decisions that hurt the lender(s); and
- default and remedy conditions, which specify the circumstances under which the lender(s) can terminate the lending relationship and their rights in such circumstances.

Debt issuance and management is thus a complex operation, and we stress only a few of its key features in this section.

2.3.1 Debt Maturity, Security, and Liquidity

(a) *Collateral.* In business parlance, lenders may lend "against assets" or "against cash flow." Lending against cash flow simply means that their lending is "unsecured," that is, not backed by assets, so that the expectation of recovering money is purely based on the assessment that the borrower will be able to generate enough cash flow. Lending against assets means that the lenders are partially protected against nonrepayment of interest or principal by a pledge of assets. That is, the lenders can repossess (seize) the specified assets in case of default. Lending is then "secured."

14. For example, Biais and Malécot (1996) argue that the low protection of creditors under the 1985 French bankruptcy law (which was reformed in 1994) and the concomitant reluctance of creditors to lend long is one of the factors explaining why French firms had more short-term debt than their American or British counterparts. French bankruptcy law still offers poor protection even to secured creditors because privately-agreed-upon procedures must be overruled by the court, which by law must favor continuation and employment over other alternatives, and because the state and the employees have priority over secured creditors in the case of liquidation.

15. See, for example, Greenbaum and Thakor (1995) for details about the way loans are structured.

16. A call provision granted to the issuer is the right for the issuer to retire the issue earlier than the stated maturity. This option is valuable because if the market interest rates fall, the issuer can retire the issue and refinance at a lower rate. The issuer must, of course, pay a higher interest rate in exchange for this privilege. Conversely, a right granted to the lender to accelerate payments or the collection of the entire loan somewhat protects the lender against default to the extent that it gives him an exit option when he receives signals of an impending default.

Various assets can be pledged: accounts receivables from trade customers,[17] inventories, real estate, equipment, or the managers' personal property. Guarantees from a government or from banks (letters of credit) can also play the role of collateral.

We will see in Chapter 4 that the pledging of assets substantially increases the availability of credit, although it comes with a number of costs (transaction costs, which are substantial, as well as other costs). For this reason, a substantial fraction of commercial and industrial lending is made on a secured basis.

(b) *Trading and liquidity.* It is customary to distinguish between public and private placements. Public bonds are issued on a "primary market" either directly by the issuer or more commonly through an underwriter (securities firm, investment bank, etc.). They are then traded in a "secondary market."[18] In contrast, private placements and bank loans are usually not traded after their issuance, although there has lately been a move toward transforming the corresponding claims into "securities" (that is, claims that are widely traded), a process called "securitization."

The chief determinant of whether a claim can be easily traded in a secondary market (is "liquid") is the symmetry of information among investors about the value of the claim. Suppose that the owner of a claim has more information about its value than prospective buyers of the claim. Buyers are then concerned by the "lemons problem": while the seller may have personal reasons to sell the claim (e.g., liquidity needs), he may also sell the claim because he knows that the claim is not worth much. The buyers are accordingly distrustful, and exchange is unlikely to occur in situations of large asymmetries of information (Akerlof 1970). This theoretical view sheds light on why some claims are liquid and others are not. As we will see, public bonds are usually fairly safe from default by the borrower. There is therefore little asymmetry of information among market participants about the value of public bonds, and public bonds are quite liquid.[19] In contrast, we will see that bank loans and privately placed debt have higher probabilities of default and may involve substantial asymmetries of information between the initial lenders and the prospective buyers in a secondary market. It is therefore not surprising that the securitization of such claims has remained limited.

(c) *Maturities.* Borrowing can be short or long term. Definitions of what short and long term mean are, of course, subjective, and depend on the instrument. For instance, public bonds with maturity under five years are labeled short term and those over twelve years long term. Bank loans under one year (which constitute roughly half of the bank loans) are short term and those over one year long term.

Short-term credit includes the following three items:

Loan commitments and lines of credit granted by commercial banks to borrowers. A loan commitment specifies a maximum loan amount, the commitment's period, and the terms of the loan (a commitment fee to be paid up front, as well as possibly a fee on unused balance; and the interest rate, often a fixed markup over a market rate of interest).

Commercial paper, the only publicly traded short-term debt. Commercial paper has had a very low default rate over the last forty years; it is unsecured, although its quality is increasingly enhanced

17. Alternatively, accounts receivables may be "factored" rather than pledged. That is, they are sold at a discount from their face value to a factoring company which then collects the payments. The supplier or trade creditor then receives cash which can be used to reduce the amount of borrowing, rather than be pledged as collateral when receivables are not factored (for an examination of the similarities and differences between the roles of cash and collateral for the availability of credit, see Chapter 4).

Similarly, the value of assets stemming from commercial transactions may be enhanced by bank guarantees (bankers' acceptances or letters of credit) granted by the buyer's bank (such guarantees are, for example, often used to finance foreign trade). The supplier's bank is then willing to provide an immediate payment to the supplier for the goods delivered in exchange for the enhanced trade credit, namely, the bankers' acceptance, because the claim on the buyer has become almost riskless. (Indeed, bankers' acceptances are widely traded and their interest rate in the market tracks closely the international cost of money to borrowers, LIBOR (the London Interbank Offered Rate on Eurodollar deposits traded between banks, that is, the interest rate corresponding to almost default-free transactions).)

18. Bonds are usually traded "over the counter" (on the OTC market), that is, through bilateral exchanges via dealers rather than in a centralized exchange as in the case of major stocks.

19. Note that the important property of bonds here is not the fact that default is unlikely, but rather its implication that information about their value is fairly symmetric. Indeed, while one might believe that low default rates make bonds pretty riskless, changes in market interest rates induce important fluctuations in their price (if they are not indexed on the market rate). So, the general rule is that symmetric information about a claim makes it more liquid regardless of its riskiness.

by "backup lines of credit" from a bank. Those backup lines of credit do not guarantee repayment by the bank to the holders of commercial paper in case of borrower default, but they provide liquidity enhancement to the borrower and therefore reduce the probability of default.[20]

Trade credit, that is, borrowing from suppliers. Trade credit is an important source of short-term financing at the individual firm level. In 1991, U.S. manufacturing firms had 13.7% of their total assets in accounts receivable and 7.4% in accounts payable. Trade credit is even more significant in some other countries (the same numbers for Japan were 24% and 13%).[21] It is typically very expensive: for instance, about 80% of the U.S. firms offer their products on terms called "2-10 net 30," which means that the buyer must pay within 30 days, but receives a 2% discount if payment occurs within 10 days. The 2% price increase over the remaining 20 days corresponds to a 37.24% annual interest rate![22]

Firms in general would prefer to be granted long-term credit because short-term credit forces them to return repeatedly to their bank or to the credit market for new money and exposes them to the risk of refusal and to the necessity of selling assets at distress prices or of cutting down on their activity. On the other hand, short-term borrowing has two key benefits: first, it returns more funds to the lenders and thus facilitates financing in the first place; second, precisely because it forces firms to return occasionally to their lenders, short-term borrowing imposes more discipline on the borrowers (the theoretical underpinnings for this argument will be examined in Chapters 5 and 6).

Long-term credit corresponds to bank loan agreements and to long-term privately or publicly placed debt. Long-term credit agreements are much more elaborate than short-term ones and involve a number of covenants. This brings us to the design of loans, to which we will turn in Section 2.3.3.

2.3.2 Credit Analysis

When contemplating short-term and especially long-term lending, lenders perform a credit analysis along several directions. They analyze the borrower's financial data (capital structure, cash flow statements, liquidity, etc.). They estimate the market and liquidation values of assets. They also look at the capability and character of the entrepreneur or top management. Bankers refer to the "five Cs of credit": character and capacity (capability), capital, collateral, and coverage (the first four Cs were just described, the fifth is simply the existence of insurance against death or disability of a key person): see Section 2.7 for more details. Chapters 3–6 will analyze the role of capital, collateral, and capability and character.

Credit analyses are also performed by third parties who do not lend to the firm. Predominant among

20. The maturity of commercial paper is often lower than one month, although it can extend to nine months. This short maturity implies that it is often rolled over. A bank line of credit is basically an insurance policy for the borrower/issuer as it allows the latter to pay back the outstanding commercial paper without having to sell off assets at "fire sale" (low) prices in case adverse market conditions or bad news about the issuer make it difficult to roll over the commercial paper.

Commercial paper in practice is meant to have low credit risk. (For this reason, only 22% of the commercial paper in the United States is issued by industrial companies, financial companies accounting for the bulk of the issues.) A clear description of the mechanics of commercial paper is Chapter 22 of Stigum (1990).

21. Rajan and Zingales (1995) report accounts payable for large firms equal to 15% of assets in the United States, 11.5% in Germany, and 17% in France. See Petersen and Rajan (1997) for an in-depth study of trade credit in the United States.

More recent numbers for the United States can be found in Frank and Goyal (2003), who more generally provide evidence about broad patterns of financing activity. They report for 1998 and for 7,301 U.S. industrial firms a percentage of book value of total assets equal to 17.7% for receivables and 10.4% for account payables.

22. Several explanations have been proposed as to why trade credit is widely observed given the high cost to the buyer. Some (e.g., Smith 1987) view it as a means for the supplier to distinguish between high- and low-risk buyers, and to learn useful information for their future relationship. Others have suggested that the underlying collateral (the products shipped, if they have not yet been resold) has higher value for the supplier than for a bank, but this does not explain why the interest rate on trade credit is much larger than that on bank loans. Brennan et al. (1988) offer a price discrimination explanation for trade credit. Wilner (1994) links the higher rate of interest on trade credit with the suppliers' poor bargaining position in a renegotiation following default: because the suppliers care much about the continuation of their relationship with the buyers, they make more concessions than banks in renegotiation. Biais and Gollier (1997) argue that suppliers may have private information about the riskiness of their clients, which implies that trade credit, if extended, provides a favorable signal about the credit quality of the clients and allows the latter to get cheap complementary financing from banks, which in turn has value to the suppliers in the context of ongoing trade relationships. Finally, Burkart and Ellingsen (2004) trace the informational superiority of trade creditors over banks to the knowledge that the transfer of the input has taken place. They argue on the basis of their theoretical model that trade credit should have a short maturity as it loses its advantage when the illiquid input is transformed into liquid output.

these are rating agencies. Their main raison d'être is that credit analysis is costly and, when claimholders are dispersed (as is usually the case for a public bond), it is efficient to centralize credit analysis in a single entity (or a small number of entities). Issuers of bonds or of commercial paper, by paying fees to rating agencies for being graded, in a sense solve the collective action problem faced by prospective bondholders.[23] One may wonder why rating agencies can have any reliability given they do not put their own money into the borrowing firm and that, even worse, they are paid by the very companies that they rate, which, of course, creates a conflict of interest. The answer is that they care about their reputation for measuring and disclosing accurately the riskiness of the claim. A good rating is worth more to an issuer if the previous issues which were given the same rating by the rating agency have had a good track record. Thus a rating agency which has the reputation for not trying to please its issuing clients can actually command higher fees from them.

Ratings are based on criteria similar to those used by banks for their credit analysis. The rating agency looks at the borrower's capital, cash flow, liquidity (including the existence of resources to meet unexpected cash demands), capability, and at the firm's line of business. What they emphasize more depends on various characteristics of the issue, in particular its maturity. For example, the main focus for commercial paper (which, recall, is very short-term public debt) is the borrower's liquidity, that is, how easy it is for the borrower to come up with cash to repay the maturing commercial paper.

While there are a number of private rating agencies, the market is still dominated by the two best known, Moody's and Standard & Poor's (S&P), which suggests that reputation is a very worthwhile asset and a strong barrier to entry. Ratings are sometimes also prepared by agencies or organizations in charge of controlling the asset quality of financial intermediaries and are then employed for prudential regulation, i.e., to verify the capital adequacy of the financial intermediary.[24]

Rating agencies use grades to measure the credit worthiness of issuers and securities. For example, S&P gives the following grades (in descending order): AAA, AA, A, BBB, BB, B, CCC, CC, C (and D for a firm in default); Moody's has a very similar notation. The grade reflects an estimate of the likelihood of default. For example, the cumulative default rate over the first ten years of a bond's life was 0.1% for an AAA rated bond and 31.9% for a B rated bond in Altman's (1989) sample. It is also customary to define a coarser partition, with "investment grade securities" being those with grades above BBB, and "below investment grade securities" or "junk bonds" being the others. As an approximation, only investment grade securities are issued, so securities below investment grade are mainly downgraded investment grade securities.[25] Needless to say, ratings, while useful, are not perfect, if only because agency problems may creep into decisions of credit-rating agencies as well (for example, they may devote insufficient resources to analyzing a security issue or they may strategically delay recognizing their past mistakes).

Lastly, like bondholders, trade creditors face a collective action problem with respect to the credit analysis of borrowers. A trade borrower often faces several dispersed lenders and it may be excessively costly for each to conduct a credit analysis. Unsurprisingly, trade creditors do rely on external ratings. Besley and Osteryoung (1985) cite a survey showing that 69% of U.S. firms use credit ratings supplied by mercantile agencies when determining credit limits for their clients.

2.3.3 The Writing of Debt Agreement Covenants

As discussed in more detail in Section 2.8, covenant writing is an important step in the lending process.

23. In the past, rating agencies collected fees from investors rather than from the issuer; but this, of course, gave rise to free riding among investors.

24. For example, in the United States, the National Association of Insurance Commissioners in 1990 issued guidelines creating six quality categories, NAIC-1 through NAIC-6, for privately placed debt. Only the top two grades, NAIC-1 and NAIC-2, correspond to investment grade ratings from major rating agencies. Investments by insurance companies in privately placed debt of below NAIC-3 quality are heavily penalized. Consequently, an important source of funding for below NAIC-3 borrowers dried up almost instantaneously. See, for example, Carey et al. (1993) and Emerick and White (1992) for more details about the guidelines (known as Rule 144A) and about their impact.

25. In the United States, below investment grade securities represented less than 4% of corporate debt in 1977. Even in the aftermath of the junk bond explosion of the 1980s, only one-quarter of the 23% of corporate debt rated below investment grade had been issued as junk bonds.

Covenants can be found to various extents in bank loan agreements, in privately placed debt agreements, and in public bonds issues. Their details depend not only on the nature of the lenders, but also on the maturity and other specificities of the claim.

It is customary to distinguish between positive and negative covenants. Positive covenants stipulate actions the borrower must take, while negative covenants put restrictions on managerial decisions. I do not find this standard distinction very enlightening: a positive covenant specifying an action may be viewed as a negative covenant prohibiting the opposite action. For instance, the obligation of maintaining assets in good repair and working order, a positive covenant, can be alternatively stated as the prohibition of letting the company's assets wear and tear. We will depart from tradition by offering a taxonomy more in line with economic considerations, which suggest two rationales for covenants.

To understand the first rationale for covenants, it is useful to recall that managers and shareholders are in control of the firm as long as the covenants are not violated.[26] Managers and shareholders often have incentives to take actions that jeopardize the payment of interest and principal to lenders (we will later divide such actions into two sets). These actions redistribute wealth from lenders to managers and mainly shareholders. Note that the fact that the actions redistribute wealth per se is not a motivation for the existence of covenants. Such actions may reduce the value of debt and increase that of equity, and yet have no impact on the total value of the firm following the Modigliani-Miller logic. Tolerating such actions through the absence of covenants lowers the value of debt, but may have no overall effect:[27] to the extent that the actions are anticipated, the *ex ante* price of bonds and equity reflects the transfer that will take place *ex post*, so that total investor value (the value of debt plus that of equity) is still the same. It is only to the extent that managers and shareholders may have incentives to take actions that reduce total firm value that covenants have a role. Thus, the first role of covenants is to prevent managers and shareholders from taking value-reducing actions that could be privately optimal because they expropriate debtholders.

The second role of covenants is to define the circumstances under which different classes of claimholder (equityholders or debtholders) receive the right to intervene in management.[28] The threat of external intervention in management is best viewed as part of the incentive package offered to insiders. As Chapter 10 will show, it may be optimal to confer control rights on shareholders in good times and on debtholders in the case of mediocre performance. The transfer of control is triggered by the nonpayment of interest or principal or by a covenant violation. This yields the second rationale for the existence of covenants. Further, to the extent that shareholders and managers are hurt by a transfer of control to debtholders, the former have incentives to manipulate the (mainly financial) measures of performance defined by this type of covenant. A further set of covenants can, however, be introduced to limit such manipulations.

Thus, our taxonomy of covenants highlights two rationales. We further divide the two sets into two subsets each.

2.3.3.1 Covenants Meant to Prevent Value Reduction (The "Conflict View")

As discussed above, the divergence of preferences between shareholders and debtholders may induce the former, when they are in control, to take actions that are meant to benefit them to the detriment of the latter. They may be willing to sacrifice total value to achieve this goal. For convenience, we subdivide the actions into two subsets depending on whether they involve an increase in the riskiness of the firm's cash flow.

Actions not increasing risk. We first consider actions that reduce the value of existing debt without

26. In principle, the shareholders, perhaps through the board of directors, are in control. In practice, asymmetric information between insiders and outside shareholders introduces an important distinction between formal authority, held by shareholders, and real or effective authority, often enjoyed by managers. For more details on this idea, see Chapter 10.

27. Unless borrowers and lenders find it easier to value debt when debt is associated with a standard set of covenants.

28. This rationale in a sense is more primitive than the first one, because it explains why claims with conflicting interests are created. The possibility of redistribution among claims, and therefore the first rationale for covenants, would disappear if there were a single claim.

per se increasing the riskiness of the firm's income flow. Covenants put restrictions on *payments* to shareholders. Payments can take different forms: cash dividend,[29] share repurchase,[30] or "affiliated transactions" (in which the firm engages in loss-making transactions, e.g., through generous transfer prices, with another unit also owned by the shareholders). Excessive payments may leave the debtholders with an "empty shell."[31]

Second, covenants impose limitations on *further indebtedness*. The issuance of new debt dilutes the value of existing debt (the reader may want to check this for the simple financial structure displayed in Figure 2.1); accordingly, limits on the amount of new debt are generally set by a covenant. Dilution is particularly strong if the new debt is either secured or senior to the current debt. It is therefore not surprising that additional covenants cover new secured or senior debt: limitations on liens; positive covenants forcing the firm to pay taxes (the government often acquiring a claim senior to that of creditors in the case of unpaid taxes) or, in the United States, to contribute to the Pension Benefit Guarantee Corporation (again, the debts to the Guarantee Corporation are senior to those of creditors); and covenants restricting leases (long-term noncancelable rental agreements may acquire some seniority, e.g., one year's lease payment, over other creditors' claims).

Actions increasing risk ("asset substitution"). As mentioned earlier, shareholders, with their convex claim, benefit from increased risk taking while debtholders, with their concave claim, are hurt. Of course, and as we noted earlier, debtholders are partially protected against gambling if their claim is convertible into equity, as they can switch to equity if the firm's income becomes riskier. But most debt claims are not convertible. Covenants are then meant to protect debtholders against increases in risk. Examples include covenants prohibiting investments into new lines of business, earmarking the loan for specified purposes, or limiting the growth of the firm; and covenants requiring life or casualty insurance for key personnel or setting minimum standards of coverage against interest rate or exchange rate risk.

It is clear that such actions, whether they increase risk or not, need not reduce total value. But each has the potential of doing so. Let us give a few examples. (i) Large payments to shareholders seriously decapitalize the firm and make it more likely that the firm will face liquidity problems or that control will be transferred to debtholders in the near future (see below). This may either demotivate the managers or induce them to "gamble for resurrection" (see, for example, Dewatripont and Tirole 1994a,b), creating value losses. (ii) Unpaid taxes in general involve late payment penalties, generating a value loss for the firm. (iii) Shareholders may benefit from issuing new debt to finance a new investment with negative net present value (NPV) simply because the loss to current and diluted debtholders exceeds the NPV loss. (iv) Risk taking may create a value loss, and yet raise the value of equity.

We now turn to the second rationale for covenants.

2.3.3.2 Covenants Defining Control Rights (The "Control View")

Shift of control in the case of mediocre performance. Some financial covenants are meant to transfer control to debtholders in the case of mediocre performance. One encounters covenants linked with the firm's (long-term) solvency. These covenants are expressed both in relative and absolute value. For example, total debt cannot exceed a fraction of total assets (leverage constraint). Or the firm's net worth (an accounting measure of equity, expressed as the difference between the book value of assets and that of liabilities) must exceed some minimum level. Interestingly, covenants also require a minimum amount of liquidity, even for long-term

29. See, for example, Smith and Warner (1979) for a description of the mathematical formulae limiting dividend distribution.

30. Share repurchases are an alternative to dividend distributions. In a share repurchase, the firm buys back its own stock and thus hands money back to shareholders (there are several modalities; see, for example, Brealey and Myers (1988, pp. 359, 360) for more details).

31. Spin-offs may be a way of expropriating debtholders. An example is Marriott Corp.'s 1992 attempt to split into two companies, a service company called Marriott International and a real-estate company called Host Marriott, a smaller and riskier concern to whom all of Marriott Corp.'s debts would have been assigned. Unsurprisingly, the initial stock market reaction at the announcement of the split was a rise of 21% of Marriott's stock price; and a bondholder lawsuit for fraud quickly ensued (*Washington Post*, November 18, 1992).

loans; for instance, the firm's working capital[32] is required to exceed some minimal level. Liquidity requirements are meant to guarantee that the firm will be able to face its short-term obligations. One may wonder why so much attention is paid to liquidity measures, since the fundamental issue is always that of the firm's solvency: for, a firm that momentarily lacks money can always make the shortfall through borrowing if its solvency is not in question. In this sense, liquidity problems are always solvency problems. Yet, and as bankers well know, solvency problems are often signaled by liquidity problems. Hence, the rationale for separate covenants on minimum liquidity.

The shift of control does not quite mean that debtholders start running the firm; they may do so occasionally if the firm is bankrupt and a receiver defending their interests is put in charge of the firm, or if they swap their debt for equity. But, more often, they will exert control indirectly by threatening not to refinance or to apply the default and remedy conditions (for example, the possibility for a bank to accelerate the collection of its entire loan) when a covenant is violated.[33] They can then impose a change in corporate policy, impose new covenants, renegotiate the claims, etc.

Completing the control view. This shift-of-control mechanism is more effective if two conditions are satisfied. First, the lenders must be well-informed in order to be able to detect a covenant violation and to properly exercise the power they have in that contingency. Second, the firm should not be able to fictitiously satisfy financial covenants through accounting manipulations.

Informational covenants. The need for lenders to be informed rationalizes a new class of covenants. Among these are covenants requiring the firm to report to the lender(s) a number of variables on a regular basis, covenants specifying extensive rights of inspection of facilities and books by the lender(s), and, in the case of a bank lender, the requirement that the firm's principal checking accounts be maintained with the bank.

Covenants limiting accounting manipulations. Financial covenants, to be effective, should not be easily manipulable. To the extent that their violation transfers part of the control to debtholders, managers and shareholders have incentives to use "creative" accounting in order to satisfy the financial covenants if needed. This motivates the existence of a further class of covenants that are meant to give credence to financial covenants. First, the lender(s) and the borrower must agree on an accounting method, in general the Generally Accepted Accounting Principles (GAAP) in the United States. But GAAP still leaves a substantial discretion. Covenants are then used to reduce this discretion by limiting instruments for creative accounting. Consider, for example, measures of the firm's solvency. The firm may have an incentive to sell assets whose market price exceeds the historical or book value, in order to increase the firm's measured net worth or to decrease measured leverage (as the cash received exceeds the accounting value of the assets on the balance sheet). The real net worth or leverage is not affected by the operation, but solvency covenants may no longer be violated. Consequently, loan agreements often prohibit the sale of more than a specified fraction (10%, 15%, or more) of the assets, or else require that the proceeds be used to pay down the debt.[34]

Another concern of borrowers is that the firm's real solvency be concealed through "off-balance-sheet activities" (recall from Chapter 1 that off-balance-sheet activities were prominent in some recent scandals in Europe and the United States). In particular, some liabilities are not incurred at present and in a noncontingent way. They are then recorded "off-balance." For example, a loan commitment promised against a fee to a borrower is off balance sheet for the bank issuing the commitment. The off-balance-sheet liabilities of a nonfinancial company include, for instance, leasing arrangements, consignment stocks

32. As measured, say, by the ratio of "current assets" (assets that will normally be turned into cash within a year) to current liabilities (liabilities that will normally be repaid within a year).

33. The borrower usually has a "cure period" of a few weeks to satisfy the covenant if the latter is violated. Because the deterioration of a financial ratio may be due to a bad realization of the environment such as a temporary shortfall in earnings rather than to managerial misbehavior, it makes sense to give the firm a chance to reestablish compliance with the agreement.

34. Another reason to limit the sale of assets may be that the proceeds of the sale could be used to buy new assets or enter new activities that would increase the riskiness of the firm's income (recall the "conflict view" of the rationale for covenants).

for dealers (who repay the manufacturer from sales), or an asset sale and repurchase agreement (which is similar to a loan, as the difference between the buyback price and the selling price constitutes de facto an interest payment). While not all off-balance-sheet financing need concern lenders, some arrangements may make the income statement and/or the balance sheet look better than they really are and help de facto breach loan or bond covenants without formally violating them. Consider, for example, a lease (long-term rental agreement) set up, as is often the case, so that lease payments are small at the start and larger later on. Suppose further that the lease specifications make cancellation costly. Then the firm's net worth is overstated as the corresponding future liabilities are off balance sheet. As another illustration, consider a firm's pledge to rescue a subsidiary if the latter gets into financial distress. This contingent liability is not recorded on the balance sheet, but is quite real. Unsurprisingly, covenants attempt to limit balance-sheet manipulations by the firm.[35]

2.3.3.3 Bankruptcy Process

Covenant violation generates trouble for the borrower. So does, of course, default. In the case of default, creditors or other interested parties, if they do not choose to roll over or forgive some of their claims, may force bankruptcy.[36] We will not discuss bankruptcy procedures both for conciseness and because the laws as well as the extent of their enforcement by courts are necessarily country- and time-specific. Let us just list a few well-known points. First, creditors are compensated according to some priority rule in the case of liquidation. For example, in the United States, (1) administrative expenses of the bankruptcy process are paid first, then come (2) unpaid taxes or debts to government agencies (e.g., the Pension Benefit Guarantee Corporation), (3) some wage claims (up to some ceiling), (4) secured and senior creditors, (5) junior creditors, (6) preferred shares, and, last, (7) equityholders. Second, many bankruptcy processes do not end up with a liquidation, although the threat of liquidation is important in the renegotiation or reorganization process. Third, secured and senior creditors obviously fare better than other creditors in liquidation. In the United States, secured creditors receive about 31% of their claims, senior creditors 36%, and unsecured creditors 8% (Brealey and Myers 1988, p. 742). For overviews of the issues with the current bankruptcy laws and for some policy suggestions, we refer the reader to, for example, Aghion et al. (1992), Bebchuk (1988), and White (1989).

2.3.4 The Overall Picture: Two Dichotomies in the Credit Market

2.3.4.1 Duality on the Lending Side

Simplifying a bit, lenders can be split into two groups, depending on the concentration of claim-holdings.

Sophisticated (concentrated, well-informed) lenders, also called relationship investors, include banks and institutional investors (e.g., life insurance companies) investing in private placements. The corresponding loans are extended by one or a few lenders, who are heavily involved in the writing of the loan, the monitoring of the covenants, and the renegotiation in case of covenant violation.

Dispersed lenders include public bondholders and trade creditors. They are numerous and face a free-rider problem. That is, they individually have suboptimal incentives to invest in information collection and monitoring of the borrower.

The empirical evidence shows that claims issued to sophisticated and dispersed lenders differ in a number of respects.

(a) *Screening.* It is customary to say that sophisticated investors perform more *ex ante* monitoring (that is, more screening or more credit analysis)

35. Our rendition of the writing of covenants is, of course, not exhaustive. For example, there are covenants restricting the purchase of claims (e.g., stocks) in other companies. Such covenants have several of the rationales discussed above: preventing the firm from engaging in self-dealing transactions with related companies, avoiding asset substitution, and increasing the transparency of financial covenants, the latter rationale being related to the issue of double gearing in prudential regulation (see, for example, Chapter 3 in Dewatripont and Tirole 1994a).

36. There is some controversy over whether creditors are well-protected by bankruptcy proceedings. In the United States (where most bankruptcy filings are made voluntarily by firm managers), Chapter 11 allows managers to remain in control and to have six months to propose a reorganization plan. The resulting procedure and the possibility of modifying priorities may enable managers to impose an unfavorable renegotiation plan to some groups of creditors.

before extending a loan. We must, of course, be careful not to take this view for granted; after all, while public bondholders perform little screening themselves, their demand for bonds on the primary market depends on the assessment or the mere presence of sophisticated agents such as rating agencies and underwriters, who have their reputations at stake. Thus, such sophisticated agents may go some way toward solving the bondholders' collective action problem and perform some of the role performed by banks and institutional investors in the case of private placements.

Yet, there is a widespread feeling that banks and institutional investors receive more information and access to management than those provided to investors in public markets.[37] Also, the illiquidity of bank loans and private placements demonstrates a superiority of the sophisticated investors' information over that of other investors.

(b) *Covenants.* Debt issued to sophisticated investors involves more and tighter covenants than public debt.[38] Commercial paper has very few covenants, and its long-term counterpart, public debt, has mainly negative covenants, while for both bank and nonbank private debt, affirmative and negative covenants are common.

(c) *Seniority/security/maturity.* There is a wide range of maturities from overnight (or even sometimes intraday) loans to very-long-term borrowing such as the 1996 successful 100-year bond issue by IBM.[39] Table 2.1 reviews the average maturities for a large sample of U.S. firms.

Loan maturity varies with the types of assets that are being financed. As Hart and Moore (1989) observe, assets tend to be matched with liabilities. Long-term loans are often used for fixed-asset acquisitions (property, machinery, etc.), while short-term loans tend to be used for working capital purposes (payroll needs, inventory financing, smoothing of seasonal imbalances). Thus the maturity of loans adjusts to the durability of the underlying collateral (if any).

Table 2.1 Maturity and priority structure of fixed claims in the United States. *Source:* Barclay and Smith (1996, Table 3). Reprinted with permission from Blackwell Publishing Ltd, Oxford.

	Percentage of total fixed claims	
	Mean	Median
Maturity		
More than one year	0.69	0.80
More than two years	0.56	0.65
More than three years	0.46	0.51
More than four years	0.39	0.39
More than five years	0.32	0.28
Priority		
Capitalized leases	0.11	0.00
Secured debt	0.40	0.31
Ordinary debt	0.38	0.21
Subordinated debt	0.10	0.00

Bank debt or privately placed debt tends to be secured and senior. Public bonds are rarely secured and are sometimes subordinated. It is also customary to distinguish the two forms of debt on the basis of maturity: bank debt often has shorter maturities. While banks indeed play a major role in providing short-term credit to firms, things are in fact a bit more complex here. First, there are forms of dispersed debt, such as commercial paper and trade credit, which have a very short maturity. Second, banks and institutional investors also issue long-term credits.[40] On the whole, James (1987) reports average maturities for the United States equal to 5.6 years for bank debt, 15.3 years for nonbank private debt, and 18 years for publicly listed debt, while Light and White (1979) report an average maturity of 35 days for commercial paper.

(d) *Renegotiation in the case of covenant violation (or nonrepayment).* According to conventional wisdom as well as some evidence, the renegotiation of

37. See, for example, Emerick and White (1992), who show how borrowers with very low or no credit ratings may still be able to obtain low-interest-rate credit from sophisticated investors, which suggests the existence of superior information acquisition.

38. See Kahan and Tuckman (1993) for a comparison of covenants for privately placed debt and public bonds. See also Smith and Warner (1979) and Carey et al. (1993).

39. IBM then borrowed $850 million in 100-year bonds.

40. For example, in the United States, insurance companies have played a major role in funding less creditworthy firms through long-term credits (five- to twenty-year debt).

covenants is easier when debt is held by sophisticated investors.[41] Asquith et al. (1994) show that 80% of the U.S. companies under distress restructure their bank debt through direct renegotiation (see also Gilson et al. 1990). Relatedly, Hoshi et al. (1990, 1991) find that Japanese firms that are in a "main-bank" coalition (*keiretsu*) invest and sell more after the onset of distress.

The ease of renegotiation may be due either to the concentration of claims or to better information of investors in the case of sophisticated lenders. It may be difficult to renegotiate with many investors, although some mechanisms are designed so as to achieve coordination among dispersed investors (nomination of a bond trustee who acts on behalf of the multitude of bondholders, possibility for the firm to offer new securities in exchange for bonds in order to lower its debt obligations).

(e) *Default and liquidity.* With the (minor) exception of junk bonds, public debt (commercial paper, public bonds) is rarely defaulted.[42] As explained above, this implies that there is little asymmetry of information among investors as to their value and that it can be widely traded in financial markets. In contrast, bank loans and privately placed debts do default (or are renegotiated under the threat of liquidation) with nonnegligible probability. There is asymmetric information among investors about their value, and the corresponding claims are much less liquid than commercial paper and public bonds.

(f) *Certification.* There is some evidence that the existence of a stake of a sophisticated investor in a firm helps the firm raise complementary funding, which suggests that the stake conveys favorable information about the creditworthiness of the firm. For example, firms raise more money in an initial public offering of shares when they have bank loans (James and Weir 1991). Also related is the evidence that the announcement of a bank loan grant raises the firm's stock price (Lummer and McConnell 1989).

(g) *Issue costs.* Issue costs (transaction costs, disclosure costs) are large for commercial paper and public debt and small for bank or nonbank private debt. In particular, issuing public bonds in the United States requires the firm to disclose key financial data, which may be a major disincentive if the firm's equity is not publicly traded (and therefore few of these data are public knowledge).

2.3.4.2 Duality on the Borrowing Side

Symmetrically to lenders, borrowers can approximately be split into two groups, depending on the riskiness of the debt they issue: high-quality borrowers tend to be well-capitalized, large, and highly rated by credit-rating agencies; conversely, low-quality (risky) issuers tend to be poorly capitalized, small, and unrated by credit-rating agencies.[43]

The two types of borrower have quite different borrowing patterns, which will later figure prominently in the theoretical analysis:

- High-quality borrowers have more long-term debt. The short-term indebtedness of large firms in the United States (recall that quality and size are strongly correlated) is 13% against 29% for small firms. The corresponding numbers in Germany are 39.5% and 55.9% (Gertler and Gilchrist 1994).
- High-quality borrowers can more easily obtain a loan commitment from a bank (Avery and Berger 1991) or issue commercial paper.[44] For this reason and the previous one, they manage their liquidity needs more easily than risky borrowers.
- High-quality borrowers can borrow (long) by issuing public debt while risky borrowers cannot. Risky borrowers must borrow from sophisticated investors.
- Unsurprisingly in view of the previous observations, high-quality borrowers suffer little and hardly reduce their investments, if at all, during a credit

41. Note that the ease of renegotiation is a mixed blessing. On the one hand, renegotiation enhances the efficiency of *ex post* outcomes; for example, it can prevent liquidation in situations in which continuation is socially optimal. On the other hand, it weakens the power of *ex ante* incentives. The firm is less concerned about the possibility of a covenant violation and the concomitant threat if it knows that the covenants will be renegotiated. That is, the prospect of renegotiation reduces discipline. For more on this, see Burkart et al. (1996), as well as the discussion of the soft budget constraint in Section 5.5.

42. For example, Stigum (1990, p. 1037) observed that only five issuers of commercial paper had defaulted in the United States during a period of fifteen years.

43. Indeed, "fewer that 25 of the over 400 industrial U.S. companies rated investment-grade by Standard & Poor's Corporation had total assets of less than $500 million as of year-end 1991" (Emerick and White 1992).

44. Commercial paper, which, recall, is unsecured short-term public debt, is mainly issued by firms with AAA or AA credit ratings.

crunch. A credit crunch is triggered by a decrease in banks' and other intermediaries' loanable funds (either because of a decrease in the intermediaries' capitalization or because of a tightening of prudential regulation or of monetary policy). Because risky borrowers are dependent on such funds, they are substantially hurt by a credit crunch. Also, bank loans to small manufacturing firms fall relative to bank loans to large firms when "money is tight" (Gertler and Gilchrist 1993; Oliner and Rudebusch 1993).

- The restrictiveness of loan covenants is inversely related to the credit quality of the borrower (Carey et al. 1993). Small borrowers also post more collateral than high-grade borrowers (Berger and Udell 1990).

2.4 Equity Instruments

Our treatment of equity financing will be a bit briefer than that of debt financing since we have already covered some of the material in Sections 1.4 and 1.5 on active monitoring by large shareholders and takeovers, respectively. We here emphasize the life cycle of equity financing from start-up and alliance financing to the initial public offering (IPO) or sale, and from there on to seasoned equity offerings. On the equity side, one central theme is, as in the case of debt, the role of delegated monitoring in alleviating the hazards attached to dispersed ownership. Since we have already reviewed the role of large shareholders, boards, and the market for corporate control in Chapter 1, we here focus on that of venture capitalists and alliance partners as illustrations of equity financing in the early stages of a firm's life (another important form of private equity with covenants with regards to the exit mechanism that are reminiscent of those for venture capital is shareholder agreements, including joint ventures[45]). We then discuss the mechanisms for issuing equity in Section 2.4.2.

2.4.1 Privately Held Equity and Sophisticated Investors: The Case of Start-up Financing

As in the case of debt, companies may need to sell their equity to some large, sophisticated investor. Three prominent classes of such investors in the case of privately held companies are venture capitalists, large customers, and leveraged buyout (LBO) specialists. As a rule of thumb, venture capitalists (venture capital partnerships, investment institutions, or wealthy individuals) and large customers provide finance for young, high-risk firms, while LBOs often concern mature firms with rather predictable cash flows. While LBO entities are highly leveraged and venture capital start-ups carry little or no debt, venture capital and LBO deals have several features in common, including high-intensity monitoring by concentrated outside equity holdings and high-powered incentives (small cash salary and substantial equity holding) for insiders. We discussed LBOs in the context of takeovers (see Section 1.5), and, not to repeat ourselves, we here focus on venture capital and large customer financing.

2.4.1.1 Venture Capital

Venture capital is used to finance start-up companies, often in high-tech industries (software, biotechnology. For instance, Apple, Compaq, Genentech, Google, Intel, Lotus, and Microsoft initially received venture capital), but also in other industries (for example, Federal Express and People Express started with venture capital). Further, venture capitalists specialize in highly risky projects (they fail to recoup their investments in many of the selected firms, but make spectacular profits on a few). Venture capitalists take concentrated equity positions[46] in the company they finance as well as seats on the board of directors. They carefully structure deals and monitor the firm. They also bring expertise and industry contacts.

(a) *Structure of deals.*[47] Like sophisticated creditors (see Section 2.3.3), venture capitalists devote much attention to the structure of deals. Screening

45. See Chemla et al. (2004).

46. In the case of a venture capital partnership, the lead venture capitalist or general partner (who performs most of the monitoring) has an average equity stake of 19% while limited partners have an average equity stake of 15%.

Our discussion of venture capital focuses on the American environment. For a discussion of the financing of high-tech start-ups in Europe, see Adam and Farber (1994).

47. For more on deal writing, see Gompers (1995), Case 9-288-014 of the Harvard Business School (1987), and Sahlman (1990). The reader will find much interesting evidence on venture capital contracts in Gompers and Lerner (1999, 2001) and Lerner (2000).

of firms is intense (a tiny fraction of proposals received are funded), and conditions imposed on firms are drastic. Venture capital deals usually include:

- A very detailed outline of the various stages of financing (e.g., seed investment, prototype testing, early development, growth stage, etc.). At each stage the firm is given just enough cash to reach the next stage.
- The right for the venture capitalist to unilaterally stop funding at any stage. That is, the venture capitalist may need no justification to stop funding. Less universally, the venture capitalist may further have a put provision, namely, a right to demand repayment of all or some of the already invested capital.[48]
- The right for the venture capitalist to demote or fire the managers if some key investment objective is not met, and a noncompete clause for key employees.
- The right to control future financing. Venture capitalists have preemptive rights to participate in new financing and have registration rights.[49]
- Often, the venture capitalist's ownership of preferred stock (often convertible into common stock), that is, of a claim senior to the manager's claim in liquidation. Eighty percent of venture capital deals in Kaplan and Strömberg's (2003) sample had the venture capitalist hold convertible preferred stocks (Sahlman (1990) and Gompers (1998) report similar findings).
- Some covenants such as the obligation to purchase life insurance for key employees.
- An exit mechanism for the venture capitalist. The expectation is that at some stage, the firm (if it has survived all previous stages) will go public and will sell shares in an IPO to other investors (e.g., pension funds, insurance companies, individual investors) and that the venture capitalist will sell part or all of her shares; or else the start-up will be purchased by a large firm.

Kaplan and Strömberg (2003) study a sample of 213 venture capital investments in the late 1990s. They document that the venture capitalists' rights (cash flow, board, voting, liquidation, and others) are often contingent on verifiable measures of financial and nonfinancial performance. An example of a financial performance measure is EBIT (earnings before interest and taxes). Nonfinancial performance measures include patent grants (or, for a pharmaceutical product, Federal Drug Administration approval), actions to be taken, or the founder remaining in the firm. Following on a good performance, the entrepreneur retains or obtains more control rights and the venture capitalist may then content himself with cash-flow rights. Conversely, a poor performance may lead to a double penalty for the entrepreneur: her financial stake in the start-up depreciates and the venture capitalist retains his control rights or acquires new ones. Selecting a subsample of 67 companies, Kaplan and Strömberg (2004) further show that, in more risky companies (entrepreneurs who are inexperienced or have failed in the past, companies whose operations are harder to observe, etc.), venture capitalists receive more control rights, have a greater ability to liquidate upon poor performance, entrepreneurs receive more contingent compensation, and financing in a given round is more contingent.

(b) *Certification and reputational capital.* Venture capitalists care about their reputational capital for (at least) two reasons (see Barry et al. 1990; Sahlman 1990; Megginson and Weiss 1991). First, a number of other parties—such as limited partners, input suppliers, providers of later-stage financing—piggyback on the venture capitalist's monitoring of the firm. A reputation for careful monitoring thus enhances the prospects of the venture. Second, if the start-up undergoes an IPO, the venture capitalist's good reputation (as in the case of a bank loan, see Section 2.3.4.1) reduces the underpricing of the firm's share at the IPO. (As one would expect, underpricing is particularly low if the venture capitalist keeps an equity position beyond the IPO to signal the quality of the new issue.) These two benefits for the firm from the venture capitalist's good reputation enable the latter to obtain a better deal from the borrower.

(c) *Comparison with sophisticated debtholders.* Debt financing is not an attractive alternative for the types of firm usually financed by venture capital.

48. Bank loan agreements usually allow the bank to collect the entire loan, that is, to accelerate its payment, only if certain covenants are violated.

49. In contrast, bank loan agreements mainly limit dilution of debt through issuance of equal priority or more senior debt (see Section 2.3.3).

First, ideas are not good collateral (recall that debt financing is often secured). Second, many such firms do not generate positive cash flows for quite a while and any short-term debt obligation could lead the firm into bankruptcy. Accordingly, such firms resort to equity financing. It is nonetheless interesting to compare the two types of financing. Venture capital deals combine several features of debt contracts with sophisticated creditors (high-intensity screening and monitoring, careful attention to the timing of funding, some control over future financing, seniority of claims, some covenants, certification) with the usual prerogatives of equity (such as a fuller right to control financing or the right to demote or fire managers). Simplifying a bit, venture capital deals involve more control rights for the financier and fewer covenants than private debt agreements.

2.4.1.2 Alliance with a Large Customer

For R&D firms, contracting with a large customer offers an alternative to venture capital financing. Indeed, research alliances surpassed public offerings in the 1990s as the dominant source of financing for biotechnology firms (Lerner and Merges 1998). A biotechnology company often enters into a research agreement with a pharmaceutical (or larger biotechnology) firm. The latter's primary role at the research stage is to provide financing; its role in production expands gradually as the project moves to the development and the marketing and sales stages. The biotechnology company is rewarded through royalties from licensing, including from the license to the partner, if the project is completed successfully.

The principal-agent relationship between the pharmaceutical company and the biotechnology unit (the R&D firm) is fraught with moral hazard. First, some dimensions are related to multitasking, as the R&D firm may juggle several research projects, including ones with other partners or on its own. Second, biotechnology companies' researchers often have academic objectives (publications requiring disclosure, reputation for a research orientation that enables the employment of postdocs, etc.) that may clash with a given project's profitability concerns. Third, reputational concerns (vis-à-vis academia or future partners) may prevent a researcher from admitting that the project is unlikely to succeed and therefore from suggesting termination.

Lerner and Malmendier (2004) study biotechnology research collaborations. Almost all such contracts in their sample specify termination rights. These may be conditional on specific events (50% of the contracts in their sample of 584 biotechnology research agreements) or at the complete discretion of the financier (39%). The financing firm may in the case of termination acquire broader licensing rights than it would have in the case of continuation. These broad licensing rights can be viewed as costly collateral pledging that both increase the income of the financier and boost the R&D firm's incentive to reach a good performance on the project.[50] Lerner and Malmendier's empirical finding is that such an assignment of termination and broad licensing rights is more likely when it is hard to specify a lead product candidate in the contract (and so entrepreneurial moral hazard is particularly important) and when the R&D firm is highly constrained financially.

2.4.2 Initial and Seasoned Public Offerings

It is customary to identify *four stages* of equity financing. In the first stage, equity is held by one or several entrepreneurs. These entrepreneurs may in a second stage raise equity capital from a small number of investors through a private placement; alternatively, they may have a privileged relationship with a bank. In a third stage (which most firms do not get to) the firm goes public in an initial public offering (IPO). Lastly, it may then conduct secondary or seasoned public offerings (SPOs). IPOs and SPOs have a strong business cycle component and are much more frequent during upswings.

2.4.2.1 The Going-Public Decision

Going public is *costly*. First, firms must supply detailed information on a regular basis to regulators and investors. This involves transaction costs as well as possibly disclosure of strategic information to product market rivals.[51] Second, the firm must pay

50. See Section 4.3.4 for the theoretical foundations of this assertion. See also Review Problem 10 for a modeling of some of the arguments.

51. Yosha (1995) argues that firms with sensitive R&D information should remain private.

substantial underwriting and legal fees. In the United States, the commissions paid to investment bankers have converged in the late 1990s to 7% of the transaction for 90% of the IPOs (Chen and Ritter 2000); they are lower in other countries.[52] A company that goes public usually issues a fixed number of shares at some prespecified price. Shares are rationed if there is excess demand at the offer price. It is well documented (Ibbotson 1975; Ritter 1987) that IPOs with a preset price are underpriced in that the shares are traded on the secondary market shortly after the IPO at a premium of 15-20% on average relative to their offer price. During 1990-1998, companies going public in the United States left $27 billion on the table, a sum twice as large as the $13 billion fees paid to investment bankers (Loughran and Ritter 2002). A standard explanation for this underpricing phenomenon is the existence of a "winner's curse" in such offerings (Rock 1986).[53] Third, the insiders (entrepreneur, venture capitalist if any) have superior information about the prospects of the firm,[54] especially if the firm has low visibility and no track record. The insiders may therefore be reluctant to sell shares at a discount when they are unable to demonstrate to investors that the firm indeed has excellent prospects. Fourth, new investors often demand control rights, especially in countries with a poor enforcement of minority rights; entrepreneurs, however, may want to retain control for themselves or within the family. As a matter of fact, family firms still dominate the corporate landscape around the world (see Section 1.4).

Firms derive several *benefits* from going public. First, going public enables firms to tap new sources of finance and to enable the firm's growth. Relatedly, it enables the firm to be less reliant on financing by a single bank or a venture capitalist; by diversifying its sources of finance, it is better protected against a "holdup" by the key financier. Second, going public facilitates exit; it allows the entrepreneurs and large shareholders to diversify their portfolios (see Pagano 1993); relatedly, it enhances the liquidity of their claims (see Chapter 9). Third, going public creates a relatively objective measure of the value of assets in place, which can be used for managerial compensation purposes (see Chapter 8). Fourth, going public may help discipline managers through the channel of takeovers.[55] On the other hand, it may reduce the intensity of monitoring by creating a more dispersed ownership structure, which has costs as well as benefits (such as the promotion of officers' initiative (Burkart et al. 1997)). Lastly, the firm's listing on a stock exchange enhances name recognition; this may help the firm not only to find new investors, but also to improve its relationship with other potential stakeholders such as trading partners or creditors.

There are few empirical investigations of the decision to go public. Pagano et al. (1998), on Italian data, show that firms in industries in which other firms have a high market-to-book ratio are more likely to go public. This may be due either to the possibility that the increased availability of funds associated with public listing is more attractive to firms with high growth prospects (this reason does not seem plausible for the Italian sample, as investment and

52. Chen and Ritter analyze several factors that may be conducive to high commissions: importance of buying underwriter prestige, possibility of tacit or explicit collusion, incentive provided to the underwriter to credibly certify the issue, nonprice competition.

"Legal fees" include registration fees, taxes, fees for legal and accounting services, and so forth. See Eckbo and Masulis (1995) for an earlier review of the empirical evidence on the magnitude of those fees.

53. Suppose that some investors have superior information about the prospects of the company than others, but that they may not buy the whole issue (because of regulatory constraints, risk aversion, etc.). The less informed investors should realize that they receive more shares when the informed investors are unwilling to buy, that is, when the company's prospects are low, and that they are rationed when prospects are high. Hence, the only way to attract less informed investors is to sell shares at the discount. (The IPO underpricing is only about 4% in France, where a mechanism resembling more a standard auction without rationing is used.) The winner's curse effect seems to be weaker when the existence of a bank loan signals that prospects are high.

Interestingly, underpricing is also smaller when the offering's underwriter guarantees the proceeds from the entire issue to the company—the method of firm commitment—than when the underwriter only offers "best efforts" to place the issue. The underwriter may well "certify" the issue better in the former case than in the latter case, in which its stake is lower. On the other hand, it might be that the higher underpricing under a best-efforts contract is due to a sample selection bias—best-efforts contracts are used mainly for smaller, speculative issues (therefore prone to substantial winner's curses)—rather than to a weaker certification by the underwriter. (See, for example, Eckbo and Masulis (1992), Hanley and Ritter (1992), Loughran and Ritter (2002), Ritter (2003), and Ritter and Welch (2002) for more information on IPOs.)

54. See Chapter 6 as well as Chemmanur and Fulghieri (1999).

55. See Chapter 11. Zingales (1995) further argues that free riding by small shareholders may help extract more surplus from prospective acquirers.

profitability decrease after the IPO) or to the possibility that firms go public in hot (high-value) markets (see Section 2.5 for a discussion of market timing). A second finding is that larger companies are more likely to go public. A third finding is that, even controlling for firm characteristics and the reduction in leverage after the IPO, firms borrow from a larger number of banks and experience a reduction in the cost of bank credit after the IPO, perhaps due to the increase in transparency or to the availability of new sources of capital. Lastly, and unsurprisingly in view of the low level of investor protection in Italy,[56] the Italian stock market is much smaller relative to the size of the economy than the American one. Relatedly, the typical Italian firm going public is eight times as large and six times as old as the typical firm going public in the United States.

A few studies (e.g., Anderson and Reeb (2003) for the United States and Sraer and Thesmar (2004) for France) attempt to analyze the relative profitability of family firms. Family firms run by their founder(s) unsurprisingly tend to be very profitable. The question is more whether firms that are run by heirs or by a professional manager hired by the family who has retained control over the firm[57] do less well than widely held firms.[58] On the one hand, one might expect heirs not to be the most appropriate choice for management (indeed, the founder may want to sacrifice wealth in order for the family to keep the benefits of control). On the other hand, the founder may have superior information about prospects and may want to keep the firm private when these are excellent. Thus, even ignoring other effects, it is not clear what we should expect.

Sraer and Thesmar (2004) use a panel of 750 corporations listed on the French stock exchange from 1994 through 2000. On that stock market, two-thirds of the firms exhibit a significant family ownership; among these, almost 50% are still managed by their founder, 30% by a heir of the founder, and 20% by a professional CEO. Consistently with previous studies on U.S. data, Sraer and Thesmar find that family ownership is associated with both higher economic and market performance. Lower wages in family firms seem to explain an important part of these higher performances. Sraer and Thesmar provide evidence consistent with the fact that, because of their different time horizons, family firms have a comparative advantage in enforcing implicit insurance contracts with their labor force. A surprising fact is that heir-managed firms do as well (in terms of return on equity or return on assets) as firms run by founders or by professional managers, and better than widely held corporations. As Sraer and Thesmar note, though, there are potential biases stemming from both the impact (alluded to above) of private information on the decision to go public and from the fact that badly managed heir-controlled firms tend to disappear or else surrender control under financial hardship.[59]

2.4.2.2 The Equity Issue Process and the Role of Underwriters

There are several flotation methods.[60] The most common way of raising equity in the United States is to use an underwriter. The underwriter may guarantee the proceeds of the shares in case of undersubscription; the underwriter can then sell the unsold shares at a lower, but not at a higher, price than the price stated in the public offering. This is the "firm commitment" contract institution. The risk borne by the underwriter is limited, though, if, as is often the case, the price is fixed shortly before the offering. By contrast, under a "best efforts" contract, the underwriter does not bear the risk of offer failure; and the offer is withdrawn if a minimum sales level is not

56. An indicator of the poor investor protection in Italy is the very high premium attached to shares with voting rights relative to shares with the same cash-flow rights but no voting rights (see Zingales 1994).

57. For example, among automobile manufacturers, Peugeot has been managed by heirs, and Fiat and BMW by professional managers.

58. In Burkart et al.'s (2003) theoretical model, a founder chooses between selling the firm, in which case it becomes widely held and is run by a professional manager, and keeping control over it, which gives the founder the option between a professional manager and a heir to run the firm. They assume that heirs are less competent than professional managers and argue that transforming the firm into a widely held company is optimal when the legal protection is high. With lower investor protection, ownership concentration is called for. Heir-managed firms, which avoid a separation of ownership and control, arise in their model when investor protection is very poor.

59. Looking for such biases, they nonetheless argue that their approach may actually underestimate the performance of heir-controlled firms relative to widely held firms, as heir-controlled firms are performing better than all other firms one year before returning private.

60. See, for example, Eckbo and Masulis (1995) and Hanley and Ritter (1992) for more extensive discussion of flotation methods.

reached within a specified amount of time. In the 1980s, firm commitment issues accounted for the bulk of SPOs of common stock in the United States, and for about 60% of IPOs. The remaining 40% of IPOs, corresponding mainly to smaller, more speculative issuers, were conducted under best-efforts contracts (Ritter 1987).

Underwriters often play the dual role of stock analysts. They subsequently issue recommendations to investors regarding the value of the securities that they have helped float.[61] Indeed, the underwriter most often implicitly commits to provide analyst coverage in the aftermarket. Conversely, even "independent" or "nonaffiliated" analysts, who have not underwritten the specific security that they are assessing (or other securities issued by the firm), may later on assist with other public offerings.[62] There is a widespread feeling that this dual role creates a conflict of interest, so that analysts have incentives to issue positive recommendations so as to please issuers and obtain future underwriting contracts.[63] In the United States, a settlement between regulators and major brokerage firms made the latter pay a fine of $1.4 billion for biased and misleading recommendations. This incentive to please issuers must be traded off against that to maintain a reputation for reliable assessments. Research has been investigating the differentials in conflict of interest.[64]

There are other ways of issuing equity, such as private placements and direct issues. A potentially important alternative to tapping new investors is to issue shares to existing shareholders through the institution of rights offers. Indeed, in North America and in Europe, existing shareholders have by law the first right of refusal to purchase a new issue of common stock. A rights offer consists in offering shares first to existing shareholders, often at a 15-20% discount under the current market price. Rights offers have become rare in the United States, but they are more common in Europe and in Japan.

Still another way of issuing equity is to transform other securities (as in the case of an equity for debt swap) or cash into equity, or to issue securities that can later be converted into equity (convertible debt, warrants, stock options). Employee stock ownership and direct reinvestment plans automatically transform employee compensation and shareholder dividends, respectively, into shares. As noted by Eckbo and Masulis (1995) in the United States, such schemes may have substituted for rights offers.

2.5 Financing Patterns

This section documents firms' financing patterns. Firms finance operating expenditures and investments in roughly two ways: (a) *retentions*, which we define as the difference between post-tax income and total payments to investors. Total payments to investors include payouts to shareholders (dividends, share repurchases), and payments to creditors (principal and interest) and to other securityholders; and (b) *return to the capital market*, that is, the issuing of new shares and bonds and the securing of new loans or trade credit.

Chapters 5 and 6 will stress the risk inherent to capital market refinancing. Unless the firm draws on a previously-contracted-for credit line or more generally is able to use some already secured source of financing, the refinancing process is confronted with investors' reluctance to lend funds whose proceeds they will imperfectly appropriate. Refinancing thus exposes the firm to the risk of being unable to

61. In the United States, they must wait 25 days to issue such recommendations.

62. While underwriters have an incumbency advantage for future offerings, a nonnegligible fraction of issuers do switch underwriters. Krigman et al. (2001), on a U.S. sample in the mid 1990s, find that 30% of the firms completing a secondary equity offering within three years after their IPO switched lead underwriter. Noting that most of the switchers do not report a dissatisfaction with their IPO underwriter, they suggest two possible explanations for this phenomenon. First, firms that started with less-well-known underwriters may "graduate" to higher-reputation ones. Second, they may "buy" additional analyst coverage from the new lead underwriter.

63. Much of the research builds upon information supplied by the company's management. The brokerage firms' revenue from providing advice to institutional investors and others is indirect. First, they receive money from future investment banking contracts with companies that are covered. Second, brokerage firms receive trading commissions from institutional investors, who if they own such shares in a company do not want the brokerage firm to publicly issue a "sell" recommendation.

64. Michaely and Womack (1999), on a sample of 1990-1991 U.S. IPOs, find that lead underwriters issue more optimistic recommendations and that the market reacts less to their recommendations. Bradley et al. (2004), on a "bubble period" sample of 1997-1998 U.S. IPOs, do not find any difference in market reaction between affiliated and nonaffiliated analysts, which they interpret as evidence that affiliated analysts have superior information or that nonaffiliated analysts are also very eager to please the company.

Table 2.2 Average financing of nonfinancial enterprises, as a percentage of total financing sources, 1970–1985. *Source:* Mayer (1990).

	Canada	Finland	France	Germany	Italy	Japan	U.K.	U.S.
Retentions	54.2	42.1	44.1	55.2	38.5	33.7	72.0	66.9
Capital transfers	0.0	0.1	1.4	6.7	5.7	0.0	2.9	0.0
Short-term securities	1.4	2.5	0.0	0.0	0.1	n.a.	2.3	1.4
Loans	12.8	27.2	41.5	21.1	38.6	40.7	21.4	23.1
Trade credit	8.6	17.2	4.7	2.2	0.0	18.3	2.8	8.4
Bonds	6.1	1.8	2.3	0.7	2.4	3.1	0.8	9.7
Shares	11.9	5.6	10.6	2.1	10.8	3.5	4.9	0.8
Other	4.1	6.9	0.0	11.9	1.6	0.7	2.2	−6.1
Statistical adjustment	0.8	−3.5	−4.7	0.0	2.3	n.a.	−9.4	−4.1

finance positive net present value (NPV) continuation projects or growth prospects.[65]

The section is organized as follows. Section 2.5.1 documents sources of finance. Section 2.5.2 discusses some key theoretical principles and empirical findings relative to payout policies, or equivalently retentions. Finally, Section 2.5.3 studies seasoned equity and debt offerings.

2.5.1 Sources of Corporate Finance

Several studies (see, in particular, Borio 1990; Corbett and Jenkinson 1994; Eckbo and Masulis 1995; Kojima 1994; Kotaro 1995; Mayer 1988; Rajan and Zingales 1995, 2003) have documented the sources of finance in different countries. Figure 2.4 and Table 2.2 illustrate some typical findings for the 1980s, due to Mayer (1988, 1990).

In all countries, internal financing (retained earnings) constitutes the dominant source of finance. Bank loans usually provide the bulk of external financing, well ahead of new equity issues, which account for a small fraction of new financing in all major OECD countries.[66] One difference among countries is the role of bond financing. Bond markets play a minor role except in North America.[67]

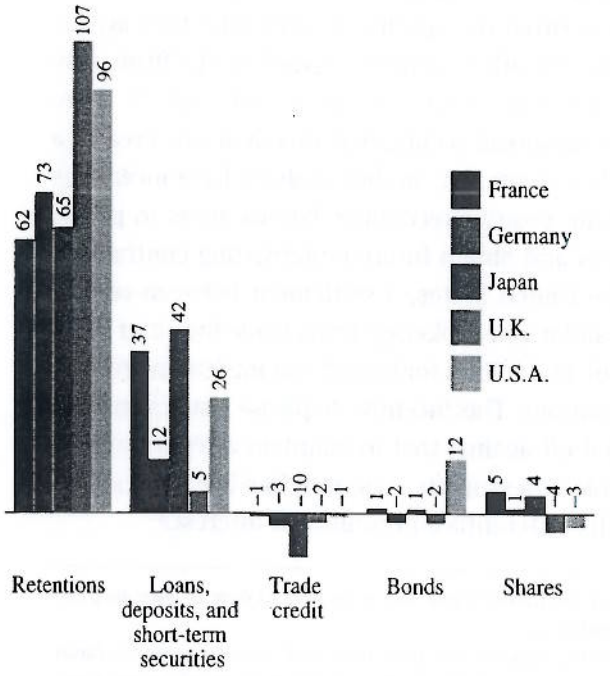


Figure 2.4 Reprinted from *European Economic Review*, Volume 32, C. Mayer, New issues in corporate finance, pp. 1167-1189, Copyright (1988), with permission from Elsevier.

The 1980s have even witnessed net retirements of equity in the United States. This does not mean that the volume of equity issues was negligible relative to that of debt issues. Indeed, Rajan and Zingales (1995) report that, in their sample of U.S. firms and for the 1984–1986 period, equity issuance amounted to 65% of external financing; equity reduction, though, accounted for 68% of external financing, and so the net equity issuance was negative and basically all external financing was debt

65. As will be discussed in Chapter 5, this agency-based feature is absent in the classic Arrow-Debreu competitive equilibrium model, which assumes that firms' income is fully pledgeable to investors and so firms incur no cost when relying solely on refinancing in the capital market when needed.

66. These numbers are, of course, net, aggregate numbers. They hide substantial differences among firms; for example, equity financing may be important for start-up firms.

67. Although large European firms now have access to Eurobonds and syndicated bank loans. See also Table 2.5 below, in which bonds represent the bulk of the "Securities other than stocks" category.

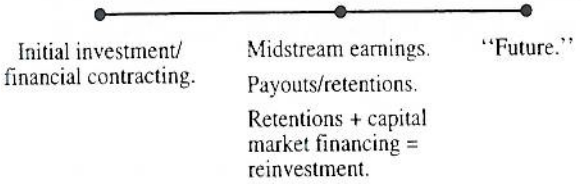


Figure 2.5

financing (primarily long-term debt issuance minus long-term debt reduction, as net short-term debt issuances were negligible).[68] The U.S. picture for the period differs a little from that for other countries over the same period. There was no equity reduction in Japan and almost none in the United Kingdom; furthermore, net equity issuance accounted for 23% and 68% of external financing in these two countries (in which external financing formed 33% and 16% of total financing, respectively). More recent data confirm the relatively minor role of equity issues in capital formation. Rajan and Zingales (2003) report that the fraction of gross fixed-capital formation raised via equity in 1999 was 12% in the United States, 9% in the United Kingdom and France, 8% in Japan, and 6% in Germany.[69]

These data should not, of course, lead us to naively overemphasize the role of "internal" financing. After all, "retentions" are cash that shareholders consent to leave in the firm for the latter to reinvest, while "equity issuances" are cash that shareholders also give to the firm for reinvestment purposes. Either way, and in a first analysis, this is cash handed over by shareholders to the firm. The difference between the two sources of finance will therefore need to be investigated in the book (see, in particular, the various discussions of the sensitivities of investment to cash flow).

2.5.2 Payout Policy and Leverage

As discussed above, there are two broad sources of financing: retentions and new securities' issues (or new loans). Because new securities' issues are hard or costly to arrange, retentions play an important role (Section 2.5.1). Yet, investors expect dividends (or share repurchases), principal, and interest, and so there is a tradeoff between retaining earnings within the firm so as to achieve continuation and growth and the need to attract investors by promising payouts to shareholders and debt repayment to creditors.[70]

Table 2.3

	Firm should retain more of its earnings if	Firm should pay out more of its earnings if
growth opportunities are	high	low
correlation of date-1 and date-2 profitabilities is	high	low
financial constraint at date 0 is	weak	tight
earnings are	small	large

To study the two key issues related to total payments to investors (payouts and debt repayments), namely, their *level* (how much?) and *structure* (what kind?), it is convenient to envision the simplified timeline in Figure 2.5 for the firm's life cycle.

The tradeoff we just alluded to refers to the tradeoff at the initial stage, "stage 0," at which the firm aims at attracting funds in sufficient quantity without jeopardizing its liquidity position midstream, at "stage 1" (more generally, the tradeoff would arise at each refinancing stage).

(a) *Payment level.* How much of the midstream earnings should be returned to investors? Intuition (to be confirmed in subsequent chapters) suggests some determinants of the payout level: see Table 2.3.

The evidence seems largely consistent with the predictions of Table 2.3. A caveat, though: the evidence presented below is incomplete. In particular, while the predictions refer to the total payment (dividend/share repurchases + principal and interest + other payments to investors), some of the evidence refers only to the dividend or the debt component of the payment. Because the determinant in question may also affect the structure of payments (e.g., the

68. External financing over the period was typically small: computed as the ratio of the net external financing to the sum of cash flow from operations and net external financing, it amounted to 14% over 1984-1986.

69. These refer to funds raised through both initial equity offerings and seasoned equity issues.

70. See Allen and Michaely (2004) for an exhaustive survey of corporate payout policies.

debt/equity ratio), it might be that the other component(s) move in the other direction.

Growth opportunities. Given the difficulties associated with returning to the capital market, the firm should pay out less when midstream reinvestment needs are high.

There is indeed much evidence that growth opportunities[71] are correlated with a lower dividend distribution (Fama and French 2001) and a lower leverage (Myers 1984).

Serial correlation of profits. The serial correlation of profits is related to growth opportunities, since, if high profits midstream are a signal of persistently high demand or low product-market competition and therefore of high future profitability, it may make sense not to distribute them and to reinvest in the firm (Poterba 1988).

Financial constraints. Recall the tradeoff between pleasing investors through high payments and promoting the firm's long-term growth through retentions. Financially constrained firms must try harder to attract funds and therefore must increase their payment ratio. There is indeed evidence that financially unconstrained firms take on low debt burdens (Hubbard 1998).

Earnings size. Intuitively, firms with low earnings midstream, controlling for growth opportunities, should distribute less than those with high earnings since a lower payment-to-earnings ratio is required in order to achieve a given level of retentions. This theoretical prediction may be less compelling than the others, though, since firms with low profits may also be financially constrained, which as we indicated above would suggest high payouts, an effect that would be further amplified by a serial correlation of profits.

The list in Table 2.3 is, of course, incomplete. For example, the derived payment policy may depend on the extent of date-0 moral hazard, as, for example, when the midstream earnings are sensitive to date-0 managerial choices. A policy of reinvesting a sizeable fraction of the profits provides management with an incentive to boost these earnings. That is, a lower payment ratio in the case of high earnings reduces moral hazard. Thus, the sensitivity of retentions to earning should increase when date-0 moral hazard increases (see Section 5.5). In the same vein, large payouts may not be advisable when management can easily reinvest earnings as they accrue and thereby hide them temporarily from investors. Lower payment ratios then incentivize management to recognize the earnings. Relatedly, firms may have an easier time secretly reinvesting money when cash flows are high (see Dow et al. 2003; Philippon 2003).

71. Empirically, growth opportunities are often proxied by the ratio of market value of assets to book value of assets.

Table 2.4 Leverage in different industries. Measures of corporate net worth by industry in the United States, 1985.

Industry	Ratio of net worth to total assets	Ratio of debt to equity
All industries	0.32	2.11
Agriculture, forestry, and fishing	0.32	2.12
Mining	0.45	1.21
Construction	0.28	2.52
Manufacturing	0.45	1.20
Transportation and public utilities	0.40	1.50
Wholesale and retail trade	0.29	2.49
Services	0.31	2.25
Finance, insurance, and real estate	0.26	2.90
Commercial banks	0.08	11.00
Savings banks[1]	0.04	28.00

Source: U.S. Internal Revenue Service,;White (1991).
1. Mutual savings banks plus savings and loan associations.

(b) *Payment structure: the determinants of financial structure.* So far, we have discussed only total payment to investors. Should this payment take the form of a fixed, predetermined payment to debtholders or a more flexible payout to shareholders? This raises the question of the firm's desired financial structure, to which we now turn our attention.

We have seen that some firms (financed by venture capital) do not contract debt liabilities. In contrast, others, following an LBO, may have debt-equity ratios of 10 or 20. Some publicly traded companies have similarly high debt-equity ratios because of the

Table 2.5 International comparison of financial structures.

	France	Germany	U.K.	Italy	U.S.	Japan
Securities other than stocks	7.3	2.3	10.6	2.3	15.6	8.0
Credit	24.3	43.2	30.7	32.1	10.0	39.5
short term	6.7	12.2	—	—	—	—
long term	17.5	31.0	—	—	—	—
Stocks	52.9	40.7	53.0	49.4	45.6	28.0
listed	17.1	—	—	—	—	—
nonlisted	30.8	—	—	—	—	—
Trade credit	15.5	8.2	5.7	12.5	8.0	17.9

Source: David Thesmar, personal communication. Table built from Eurostat, Federal Reserve Board, Bank of Japan; year 2002; fraction of total liabilities of nonfinancial corporations; fractions may not add to 100 since some lines have been omitted, to ease readability. "Securities other than stocks" are basically bonds. Also "Trade credit" is not netted out with trade credit on the other side of the balance sheet.

low cash-flow risk: for instance, banks[72] and, before the deregulation of the 1980s and especially the 1990s, public utilities (such as telephone, electricity, gas companies).[73] Bradley et al. (1984) find that U.S. telecommunications and gas and electricity companies had ratios of book value of long-term debt to book value of long-term debt plus market value of equity of 51.5% and 53%, respectively (as opposed to 29.1% for an average contemporary U.S. firm).

Measures of leverage vary substantially across studies for several reasons. For example, comprehensive samples include large numbers of small firms, which presumably are more levered than larger ones; and so leverage ratios are higher than in studies focusing on smaller samples (for example, that of listed firms). For the same reason, studies that report nonweighted means are likely to report higher leverage than those that compute weighted averages. Another reason why statistics vary widely is that studies differ in the period they cover and that leverage is time-dependent (for instance, it depends on the business cycle). Table 2.4 (due to White, who reports on a very large, nonweighted 1985 sample of U.S. firms) depicts the ratio of equity over debt plus equity in the left column and the ratio of debt to equity in the right column; a typical debt-equity ratio in this sample lies around 2.

The aggregate market-based average ratio has been remarkably stable in the United States at around 0.32 over the past half-century in the United States (Frank and Goyal 2004).

Table 2.5 (based on national accounts, and therefore weighting firms by their size, leading to lower measures of leverage) provides more recent data for France, Germany, and the United Kingdom.

Key findings about the empirical determinants of leverage are:[74]

(i) Firms that are safe (e.g., utilities before the deregulation), produce steady cash flows, and have easily redeployable assets that they can pledge as collateral (e.g., aircraft for airline companies or real estate) can afford high debt-equity ratios.

(ii) In contrast, risky firms, firms with little current cash flows, and firms with intangible assets (e.g., with substantial R&D and advertising) tend to

72. Banks are fairly riskless both because of tight prudential regulation (which, incidentally, offers a number of analogies with the analysis of covenants in Section 2.3.3) and because of deposit insurance and of the expectation that formally uninsured deposits will benefit from an implicit governmental guarantee in the case of distress. Currently, international standards impose, among other requirements, a minimum ratio of equity over (risk-weighted) assets of 8% for banks.

73. Anglo-Saxon utilities used to be regulated under the so-called cost-of-service or rate-of-return regulation, which by and large guaranteed them a safe return. The introduction of higher-powered schemes (price caps, sliding scale plans, etc.) in the 1990s made them riskier, and leverage accordingly decreased.

Regulated utilities traditionally faced little upside and especially downside risk, as regulators allowed rate increases when the utility performed poorly and strove to capture the rent through rate cuts or other means if the firm became very profitable. One substantial difference with LBOs, however, is that managerial incentives were weak. In the United States, top managers of utilities received definitely fewer bonuses and stock options than their nonregulated counterparts (see, in particular, Joskow et al. 1993), who, in turn and as we saw in Chapter 1, have much weaker incentives than managers in LBOs.

74. See Allen et al. (2005), Frank and Goyal (2004), Harris and Raviv (1992), Masulis (1988), and Titman and Wessels (1988).

have low leverage. Companies whose value consists largely of intangible growth options (high market-to-book ratios and heavy R&D spending) have significantly lower leverage ratios than companies whose value is represented primarily by tangible assets.

Remark (share repurchases and dividends). Equity payouts come in two forms: dividends and share repurchases. Share repurchases have grown substantially over the years. In particular, distributions associated with open market repurchase programs in the United States grew from $15.4 billion to $113 billion between 1985 and 1996 while dividends grew from $67.6 billion to $141.7 billion (Jagannathan et al. 2000).

In a frictionless world, the choice between the two would be neutral. It is therefore not immediately clear why firms pay so much attention to the split. Lintner (1956) postulated that dividends distribute "permanent cash flows" while repurchases distribute "temporary ones." This postulate seems more driven by the desire to account for the observed smoothness of dividends and the related observation that repurchases are very volatile (large during booms and low during recessions) than by theoretical considerations.

The world, however, is not frictionless. Taxes may differentiate the two.[75] Also, employee stock options (which, recall from Chapter 1, grew substantially in the last two decades) do not perfectly adjust for the distribution of dividends; that is, the value of options decreases when the stock goes ex dividend, which creates an incentive for management to push for share purchases (Jolls 1998).

(c) *Sensitivity of investment to cash flow.* A number of papers relate cash flow and investment. A standard finding is that firms with more cash on hand and less debt invest more, controlling for investment opportunities.[76] There are questions about what this relationship means. Were the firms at the initial financing stage ("stage 0" in our simplified timeline), more cash would ease financial constraints and therefore would indeed boost investment, as we will see in the next chapter. However, sensitivity of investment to cash flow is demonstrated in samples of ongoing concerns ("stage 1" in the timeline). One must then ask, why isn't any extra cash simply returned to investors? It may be, as we noted above, that the retention of some of the extra cash rewards management for good performance.

An alternative hypothesis is that corporate governance is far from perfect. A few papers indeed point in this direction. Blanchard et al. (1994) study large cash windfalls from legal settlements unrelated to the firm's ongoing line of business. They show that firms' acquisitions increase with these cash windfalls. Lamont (1997) shows that shocks to the price of crude oil has a substantial impact on nonoil investments of companies with an oil stake. Clearly, managers are not responsible for the oil price increase and therefore are not being rewarded for the extra cash flow.[77] Lastly, Philippon (2003) finds that investments of firms with bad governance are more cyclical than those of firms with good governance.

A more controversial finding, due to Fazzari et al. (1988), is that firms that are more financially constrained exhibit a higher sensitivity of investment to cash flow. The theory is actually rather ambiguous as to whether this should be the case.[78] Using a different approach to measuring financial constraints, Kaplan and Zingales (1997) in contrast find that *less* financially constrained firms exhibit a greater sensitivity of investment to cash flow.

2.5.3 Seasoned Financing

Let us now turn to the second broad source of refinancing: firms can conduct seasoned equity offerings (SEOs), issue new bonds, or borrow from banks.

(a) *Informational impact of SPOs and borrowing.* A well-established fact is the average permanent

75. See, for example, Jagannathan et al. (2000) for the United States. Dividends and share repurchases are treated the same at the corporate level, but repurchases had a tax advantage at the individual tax level (which was reduced by the tax reform in 1986).

See Grullon and Ikenberry (2000) for an overview of what is known about stock repurchases.

76. See the surveys by Hubbard (1998) and Stein (2003), and the many references therein.

77. Unless they are being rewarded for accurately forecasting the oil price increase. But this possibility would apply only to those managers who invested more than average in oil production. In any case, the hypothesis of a poor governance in the oil industry is to be entertained in view of the independent evidence collected by Bertrand and Mullainathan (see Section 1.4).

78. See Kaplan and Zingales (1997, 2000) and Chapter 3 for the case of initial financing and Chapter 5 for the case of an ongoing concern.

Table 2.6 Impact of financing on stock price. *Source:* Eckbo and Masulis (1995).

Type of security offered	Flotation method	Type of issuer: Industrial	Type of issuer: Utility
Common stock	Firm commitment	−3.1 (216)	−0.8 (424)
	Standby rights	−1.5 (32)	−1.4 (84)
	Rights	−1.4 (26)	−0.2 (27)
Preferred stock	Firm commitment	−0.78* (14)	0.1* (249)
Convertible preferred stock	Firm commitment	−1.4 (53)	−1.4 (8)
Convertible bonds	Firm commitment	−2.0 (104)	n.a.
	Rights	−1.1 (26)	n.a.
Straight bonds	Firm commitment	−0.3* (210)	−0.13* (140)
	Rights	0.4* (11)	n.a.

Reprinted from *Handbook in Operations Research and Management Science: Finance*, Volume 9, E. Eckbo and R. Masulis, Seasoned equity offerings: a survey, Copyright (1995), with permission from Elsevier. Average two-day abnormal common stock returns and average sample size (in parenthesis) from studies of announcements of SPOs by NYSE/AMEX listed U.S. companies. Returns are weighted average by sample size of the returns reported by the respective studies (all returns *not* marked with a "*" are significantly different from 0 at the 5% level).

fall in stock price of about 3% in the wake of an announcement of a seasoned equity issue (Asquith and Mullins 1986). (The price decrease is much less pronounced for public utilities: −0.68% as opposed to −3.25% for the 1963–1980 period in the United States, according to Masulis and Korwar (1986). It is also interesting to note that there were more common stock offerings by utilities than by industrial firms during that period, even though utilities are only a small fraction of stock market capitalization. The price decrease is also smaller in Japan (see Kang and Stulz 1994).)

In contrast, the firm's stock price rises when a bank loan agreement is announced (James 1987) although the effect seems to be driven mainly by the successful renegotiation of existing bank loans (Lummer and McConnell 1989).

There is little impact of straight debt offerings on stock prices (Eckbo 1986). Table 2.6 reports Eckbo and Masulis's (1995) summary of existing evidence for industrial firms and public utilities in the United States.

Other and related stylized facts are that the stock price increases with an announcement of higher dividends, decreases with an equity for debt swap, and increases with a debt for equity swap.

(b) *Market timing.* The link between financing and the business cycle is one of the best-documented facts in corporate finance:

(i) Bank finance is countercyclical (see Bernanke et al. 1994); firms which can afford to issue public debt in economic booms often turn to banks to meet their financing requirements during recessions. The percentage of long-term bank loans that are unsecured varies inversely with business conditions.

(ii) Firms with strong balance sheets may extend more trade credit to weaker firms and issue more commercial paper in a recession.[79] Commercial paper and bank loans move in opposite directions (Kashyap et al. 1993). Loanable funds are smaller in recessions, while there is a countercyclical demand for short-term credit.[80]

(iii) Smaller and medium-sized firms, who rely more on banks, are more affected than larger firms by business cycle-related fluctuations (Gertler and Gilchrist 1994).

(iv) Equity issues are more frequent in upswings of business cycles, both in absolute terms and relative to debt issues.[81]

(v) The negative stock price reaction to common stock issues is smaller during expansions.

(vi) Equity issues are also more frequent after an increase in the firm's own stock value.

Particularly striking is equity market timing: firms issue shares at high prices and repurchase them at low prices. Conversely, firms tend to repurchase

79. See Calomiris et al.'s (1995) study of the U.S. slowdown of 1989–1992.

80. For more on the transmission mechanism, see, for example, Bernanke and Blinder (1992), Kashyap and Stein (2000), and Kashyap et al. (1993).

81. See Eckbo and Masulis (1995) for a review of the evidence. Relatedly, stock repurchases tend to follow a decline in stock prices.

shares when values are low. This is supported by both empirical evidence (see Baker and Wurgler (2002) for a survey and Baker et al. (2003)) and survey evidence (Graham and Harvey 2001). Relatedly, corporate investment and stock market values are positively correlated both in time-series and cross-section analyses; and high stock market values such as those of the late 1990s are conducive to mergers and acquisitions in which deals are for stocks rather than cash.[82]

An interesting question is why firms time the market so carefully. There are several hypotheses in this respect.[83]

Marginal productivity. Standard neoclassical economies can partly account for a correlation between high market values and high investment. Good news about the marginal productivity of capital or low interest rates (triggered, say, by large savings rates) raises the value of firms and at the same time the profitability of new investments. If, furthermore, new investments are financed through new equity issues, then there is a close relationship between market values and equity issues (see, for example, Pastor and Veronesi 2005). The relationship is likely to be weaker, though, if to finance the new investments, debt issues or retentions—perhaps associated with high current cash flows which signal high future ones—are used instead. Note that the Modigliani-Miller Theorem unfortunately does not provide much help in predicting which source of finance is tapped.

Lower adverse selection during booms. It may be the case that adverse selection is smaller during booms, as refinancing is then more likely to be driven by new investment opportunities rather than by the desire to issue overvalued shares. Choe et al. (1993) indeed show that the negative price response to seasoned common stock offerings is significantly lower during booms. So, to the extent that firms cannot issue only debt if they want to avoid the hazards associated with higher leverage ratios, issuing equity in good times may be a wise strategy.

82. See Shleifer and Vishny (2003), who argue that managers attempt to arbitrage incorrect stock market valuations.

83. This is not meant to be exhaustive. For example, the existence of abundant liquidity in good times (see Chapter 15) may encourage more investment.

Bubbles. A couple of theoretical papers show that investment through share issues is particularly profitable in high-bubble times (Olivier 2000; Ventura 2005). Such rational-bubble models thus predict a strong correlation between equity issues and high market valuations.

Irrational markets. Several authors have lately argued that managers wait for market exuberance to issue shares. Managers who know the value of their firms better than investors and are incentivized by stock options to raise the firm's shareholder value should indeed recommend equity issues during booms and equity purchases during recessions to their board and shareholders. Note that in this argument the irrationality of investors may not stem per se from their lack of knowledge of the firm's true value (unless they fail to recognize the macroeconomic pattern of correlation), but rather in their failing to understand the adverse selection they face.

Whatever the reason, market timing is likely to have permanent effects on firms' capital structure, as documented by Baker and Wurgler (2002). And it is likely to have a differentiated impact on firms (Baker et al. (2003) find empirical support for the idea that firms that are most dependent on equity—young, highly leveraged, high cash-flow volatility, low cash-flow firms—exhibit a stronger correlation between stock prices and subsequent investment).

2.6 Conclusion

The purpose of this chapter has been to give a concise overview of corporate financing. The theoretical analysis will build on a number of themes that have become evident in this chapter, namely, the key role played by information and incentives in general, and by capital, liquidity, value of collateral, and external monitoring more specifically.

Appendixes

The following two texts are rather representative of the business world's approach to loan agreements. The first describes the five Cs of credit analysis

mentioned in Section 2.3.2. The second provides a detailed description of loan covenants.

2.7 The Five Cs of Credit Analysis

The text in this section is from a Harvard Business School note on acquiring bank credit.

When asked how a banker evaluates a borrower's creditworthiness, one is likely to hear about the "five Cs of credit analysis": the character, capacity, capital, collateral, and coverage of potential borrowers. Below, we discuss what these five Cs refer to and how they are analyzed.

Character. For many bankers, character determines if a small business loan will be approved at all. The potential trouble involved in dealing with questionable characters—noncooperation with the bank, fraud, litigation, and write-offs—are a significant deterrent. The time, legal expense, and opportunity costs incurred due to a problem loan far outweigh the potential interest income derived. (This factor, however, is less important with larger companies managed by a team of individuals.)

Capacity. Capacity refers to the borrower's ability to operate the business and successfully repay the loan. An assessment of capacity is based on management experience, historical financial statements, products, market operations, and competitive position.

Capital structure. A bank draws comfort from a capital structure with sufficient equity. Equity serves as a layer of capital to draw upon in the course of operations so as to protect the bank's exposure. Bankers also view equity as an indication of the borrower's commitment to his business. They derive greater comfort from knowing that the borrower has much to lose if his business loses.

Collateral. Collateral is the bank's claim on the borrower's assets in case the business defaults on the loan or files for bankruptcy. The bank's secured interest generally gives it a priority over other creditors in claiming proceeds from liquidated assets. The bank may also require that the borrower pledge as collateral personal assets outside of the business. For bankers, collateral is security and an alternative source of repayment beyond cash flow.

Coverage. Coverage refers simply to business insurance or "key-man" insurance which is often required when management ability is concentrated in a few individuals. In the event of the death or disability of a key manager, such coverage ensures that the bank will be repaid if the business cannot meet its obligations.

2.8 Loan Covenants

The text in this section is from Zimmerman (1975).

Loan agreements are a source of confusion and misunderstanding to many bankers. Frequently, the reader of loan agreements is not aware of their objectives and limitations, and, furthermore, is bewildered by the legal jargon of the numerous qualifying clauses.

Essential to the creation of effective loan agreements are the affirmative and negative covenants, which specify what the borrower must and must not do to comply with the agreement. The thrust of this paper is to facilitate the understanding and use of covenants in loan agreements. The use of covenants will be discussed in detail following an overview of the purpose, characteristics, and basic composition of loan agreements.

Purpose of Loan Agreements

Large amounts of time, effort, and money are spent in the development and implementation of loan agreements. They provide protection and communication for the parties involved and a general stability for the loan relationship through greater understanding among the parties. Further, should the borrower have other long-term debt, the loan agreement coordinates any legal or procedural interface with the debt and its associated creditors.

Where several banks are participating in a large credit, the loan agreement specifies the rules which govern the loan administration, and the responsibilities and liabilities of each bank.

As a major objective, the lender is interested in protecting its loan and assuring timely repayment. Through the loan agreement, the bank creates a clear understanding with the borrower as to what is expected of it. In doing so, the bank establishes its control of the relationship and provides for several basic functions to effect that control.

The lender attempts to ensure regular and frequent communication with the borrower by using certain covenants in the loan agreement. The communication results in an up-to-date assessment of the borrower's financial situation and its general management philosophy.

When the bank requires that the borrower maintain certain financial ratios, it is accomplishing several objectives. On the surface these covenants provide triggers or early-warning signals of trouble, which will allow the bank to take rapid remedial action. The borrower is made aware of where the minimum performance cutoffs are. However, the banker is also helping the borrower set reasonable goals in terms of financial conditions and growth. In some cases a "growth formula" is created which states that until a specified set of financial conditions is met, the borrower may not be eligible for further debt.

All these controls—required ratios, ratio goals, required actions, and forbidden actions—may seem arbitrary or restrictive; but applied wisely, they are not. The process lets

all parties know where they stand, thus reducing the number of unknowns or uncertainties in the loan relationship.

Characteristics of Loan Agreements

When asked to describe the salient characteristics of loan agreements, most bankers will use adjectives such as "long" or "dull" or "confusing." While many agreements may be thus described, other definitions are certainly more informative.

The loan agreement is one of the most important loan documents in that it provides the basis for the entire banking relationship, establishing intents and stating expectations. It relates all the basic loan documents to one another and creates the means of control and lines of communication which are important in protecting all parties involved.

It follows that only three main courses of action are open to the bank in the event of default by the borrower. The account officer may waive, either temporarily or permanently, the condition which has been violated. This is frequently done in the case of financial ratios, although too lax an attitude in this respect can lead to a loss of control and an ineffective covenant and/or loan agreement. An alternative is for the banker to have the agreement rewritten to make it more viable. The rewrite is also a tactic used to obtain a much tighter hold over the borrower, if needed, by using as a bargaining tool the bank's legal right to call the loan. The third, and most drastic, approach for the bank is, of course, to declare the borrower in default, call the loan, and, if necessary, file suit against the borrower.

The implications of the nature of a loan agreement are extremely important. As an example, assume that a loan has been made on an unsecured basis and one covenant forbids the pledging of assets to anyone. This is obviously an attempt to maintain the strength of the bank's unsecured position in the event of liquidation. However, let us further assume that in violation of the agreement, the borrower pledges its assets to another lender. The bank certainly retains its option to call the loan, but the other lender holds the security. If the bank does call the loan, forcing liquidation, it remains an unsecured creditor vying for those assets which remain after satisfaction of the first lienholder.

The loan agreement, then, is not a substitute for security. If a loan should be secured in the absence of an agreement, then security should be taken with one. In fact, a loan agreement is not a substitute for anything. If the situation does not satisfy the five Cs of a loan decision—character, capacity, capital, conditions, and collateral—then the loan should not be made.

Composition of a Loan Agreement

There are seven basic sections of standard loan agreements, any of which may be modified, depending upon the purpose of the loan.

- *The loan.* This section describes the loan by type, size of commitment, interest rate, repayment schedule, and security taken, if any. Also specified are all participants and their roles plus terms of participation if more than one lender is involved. Any definitions of financial accounting or legal terminology to be used in the agreement are stated here.

- *Representations and warranties of borrower.* Basically, this section is an attestation to the lender that certain statements are true. For instance, the borrower may warrant that it is a corporation, that is entering into the agreement legally, that financial statements supplied to the bank are true, and that no material change has occurred since their preparation. The company may attest to the nature of its business, that it does own its assets as represented, and that it currently is not under litigation. In other words, the company reaffirms in writing all those things about its current state of existence which have been known or assumed throughout the negotiations.

- *Affirmative covenants.* In contrast to the warranties, which attest to existing fact, affirmative covenants state what action or event the borrower must cause to occur or exist in the future.

- *Negative covenants.* Negative covenants state what action or event the borrower must prevent from occurring or existing in the future.

- *Conditions of lending.* This section states that, prior to the lending of any money, all documents and notes must be in proper form, that both the borrower's and the bank's counsel must approve the entire arrangement, and that the borrower's auditor, or at least its chief financial officer, must certify current compliance with all conditions of the loan agreement.

- *Events of default.* Conditions which will be considered events of default are specifically stated. Such conditions might be delinquent payment, misrepresentation, insolvency proceedings, change in ownership, or other occurrences which could jeopardize the company's viability and/or the bank's position. All covenant violations are considered events of default, although many are designed to be used in correcting a situation rather than in calling the loan. In any event of default, timing is crucial. For instance, it may be that default does not occur until a covenant has been violated for thirty consecutive days.

- *Remedies.* The remedies section spells out what the bank may do in the event of default. The bank's rights may include several potential actions, but always include the right to accelerate payments, a term which means to call the loan. Timing is important. The borrower may have a certain amount of time to correct the default prior to the enforcement of a remedy. In a credit with several participating banks, the remedies section also defines procedures

for calling the loan. For example, the agreement may require banks representing 70% of the commitment to call the loan.

Approach to the Covenant Package

Prior to writing a set of covenants for a loan agreement it is necessary to have a systematic approach to developing them. One must ask questions ranging from an assessment of basic objectives and risks to types of protection and remedy which must be provided to ensure the successful attainment of the objectives.

Since covenants are the heart of a loan agreement, setting the objectives is a process very similar to that of defining those for the total agreement. The bank is obviously hoping to be repaid on a timely basis, but, as a secondary set of objectives, would like to maintain or improve upon the financial position, cash flow, growth progression, and general financial condition of the borrower. Once goals have been set for the mutual benefit and protection of all parties, the lender must reassess the risks involved from a point of view different from that in the initial loan decision.

Determination of Risk

No longer is the lender looking for a yes/no decision. The aim at this point is to define the risks involved and to determine their magnitude. The account officer needs to ask, What conditions or events could block the accomplishment of my objectives? In other words, Where is the loan vulnerable? Weaknesses may lie in poor cash flow, thin net worth, or other financial statements items. It may be that the industry is volatile and highly subject to strikes or public fancy. Perhaps the company is small or it has a short track record, so that much of the loan decision is based upon projections.

Whatever the risks, it is now the task of the loan agreement writer to prevent or minimize the consequences of those risks as well as possible, in a form which remains as flexible as possible.

Scope of Covenants

The lender's effort to safeguard the loan against known and unknown risks will take the form of loan covenants. In asking what triggers exist and what actions may reasonably be taken and enforced once a risk materializes, the scope of potential covenants is almost limitless. Triggers may range from financial ratios and limits on financial statement accounts to restrictions on corporate, or even management, activities.

Furthermore, methods of treating a specific item are quite flexible in order to obtain the appropriate coverage. For example, it is possible to restrict a financial statement item to a minimum or maximum of

- a fixed dollar amount;
- a dollar amount increase or decrease per time period;
- a percentage of total assets, tangible net worth, or some independent indicator;
- a percentage change per time period.

As a special case, businesses subject to seasonal variances may have the above modifications fluctuate with the peaks and troughs of the cycle to more closely approximate actual conditions.

With so many potential requirements and restrictions, however, it becomes evident that the key to an effective loan agreement is not to see how many activities or conditions can be covered: it is to obtain the most protection in the simplest, most efficient manner.

Simplicity and Efficiency

To devise a simple and efficient network of covenants, it is imperative that the writer have a thorough understanding of the company, its management, and loan-associated risk in conjunction with a realistic attitude. This combination will result in covenants which allow the borrower maximum flexibility within the constraints necessary to provide the bank maximum protection.

(1) The borrower will maintain an adequate cash flow.

(2) The borrower will maintain a ratio of cash flow to current maturities of long-term debt of 1.5 to 1 on a fiscal-year basis.

The necessity for a realistic attitude dictates that a covenant also be such that the borrower is able to comply with it and the lender is willing to enforce it. Should either of these conditions not be met, a covenant may be frequently waived, thereby losing its psychological and, perhaps, legal control.

The essence of a loan agreement covenant is that it is simple, well-defined, measurable, risk-reducing, efficient, and reasonable. In short, it is the creative development of protection in the loan situation. As an aid to the direct application of these principles, a working guide to the construction of loan agreement covenants follows.

Working Guide for Loan Agreement Covenants[84]

Functional Objectives

The key objectives are described as follows:

- *Full disclosure of information.* To make competent, ongoing lending decisions, the account office must have an intimate understanding of the borrower. Full disclosure also aids the lender in maintaining regular contact with the borrower and close control over the loan relationship.

- *Preservation of net worth.* The borrower's basic financial strength and ability to support debt and absorb downturns

84. Only the first section of the working guide is reproduced here.

lie in its net worth. The purpose of related covenants is to assure the growth and continued strength of that net worth.

• *Maintenance of asset quality.* Asset value represents two major factors of importance to the lender: earning power and liquidation value. In either case, it is to the bank's advantage to require high standards of asset quality.

• *Maintenance of adequate cash flow.* In the case of normal repayment of a loan, the lender is repaid from the borrower's cash flow. In such cases, it is imperative that the lender closely monitor the cash flow and attempt to maintain its quality.

• *Control of growth.* As a definite drain upon cash flow, working capital, fixed assets, management energies, and capital funds, excessive growth has been recognized as the cause of numerous charge-offs and bad loans in the past few years. It is obviously in the interest of both banker and borrower to maintain growth in an orderly fashion although the two parties rarely see eye to eye on this matter. The bank's objective is to reach a clear understanding with the borrower on the limits of its growth.

• *Control of management.* In any loan situation, but particularly if the loan is unsecured, the success of the total relationship depends heavily upon the borrower's management. The bank then hopes to ensure the continuing quality of management.

• *Assurance of legal existence and concept of going concern.* The purpose of devising covenants such as these is to ensure the banks of a viable entity which may produce the conditions necessary to repay its loan.

• *Provision for bank profit.* Banks lend money in return for an expected profit, and are therefore interested, not only in protecting the principal amount of the loan, but also the profit, whether it be interest, servicing income, or other.

References

Adam, M. C. and A. Farber. 1994. *Le Financement de l'Innovation Technologique: Théorie Economique et Expérience Européenne.* Paris: Presses Universitaires de France.

Aghion, P., O. Hart, and J. Moore. 1992. The economics of bankruptcy reform. *Journal of Law, Economics, & Organization* 8:523-546.

Akerlof, G. 1970. The market for "lemons": qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics* 84:488-500.

Allen, F. and R. Michaely. 2004. Payout policy. In *Corporate Finance: Handbook of the Economics of Finance* (ed. G. Constantinides, M. Harris, and R. Stulz), pp. 337-429. Amsterdam: North-Holland.

Allen, F., R. Brealey, and S. Myers. 2005. *Principles of Corporate Finance,* 8th edn. New York: McGraw-Hill.

Altman, E. 1989. Measuring bond mortality and performance. *Journal of Finance* 44:909-922.

Anderson, R. and D. Reeb. 2003. Founding-family ownership and firm performance: evidence from the S&P 500. *Journal of Finance* 58:1301-1328.

Asquith, P. and D. W. Mullins, Jr. 1986. Seasoned equity offerings. *Journal of Financial Economics* 15:61-89.

Asquith, P., R. Gertner, and D. Scharfstein. 1994. Anatomy of financial distress: an examination of junk bond issuers. *Quarterly Journal of Economics* 109:625-658.

Avery, R. and A. Berger. 1991. Loan commitments and bank risk exposure. *Journal of Banking and Finance* 15:173-192.

Baker, M. and J. Wurgler. 2002. Market timing and capital structure. *Journal of Finance* 57:1-32.

Baker, M., J. Stein, and J. Wurgler. 2003. When does the market matter? Stock prices and the investment of equity-dependent firms. *Quarterly Journal of Economics* 118: 969-1006.

Barclay, M. and C. Smith. 1996. On financial architecture: leverage, maturity and priority. *Journal of Applied Corporate Finance* 8(4):4-17.

Barry, C., C. Muscarella, J. Peavy, and M. Vetsuyens. 1990. The role of venture capital in the creation of public companies. *Journal of Financial Economics* 27:447-471.

Bebchuk, L. 1988. A new approach to corporate reorganizations. *Harvard Business Review* 101:775-804.

Berger, A. and G. Udell. 1990. Collateral, loan quality and bank risk. *Journal of Monetary Economics* 25:21-42.

Bernanke, B. and A. Blinder. 1992. The Federal Funds Rate and the channels of monetary transmission. *American Economic Review* 82:901-921.

Bernanke, B., M. Gertler, and S. Gilchrist. 1994. The financial accelerator and the Flight to quality. National Bureau of Economic Research, Working Paper 4789.

Besley, S. and J. Osteryoung. 1985. Survey of current practices in establishing trade credit limits. *Financial Review* February:70-82.

Biais, B. and C. Gollier, C. 1997. Trade credit and credit rationing. *Review of Financial Studies* 10:903-937.

Biais, B. and J. F. Malécot. 1996. Incentives and efficiency in the bankruptcy process: the case of France. The World Bank, PSD Occasional Paper 23.

Blanchard, O. J., F. Lopez-de-Silanes, and A. Shleifer. 1994. What do firms do with cash windfalls? *Journal of Financial Economics* 36:337-360.

Borio, C. 1990. Patterns of corporate finance. Bank for International Settlements, Basel, Working Paper 27.

Bradley, M., G. Jarell, and H. Kim. 1984. On the existence of an optimal capital structure: theory and evidence. *Journal of Finance* 39:857-878.

Bradley, D., B. Jordan, and J. Ritter. 2004. Analyst behavior following IPOs: the "bubble period" evidence. Mimeo, Clemson University.

Brealey, R. and S. Myers. 1988. *Principles of Corporate Finance*, 3rd edn. McGraw-Hill.

Brennan, M. and A. Thakor. 1990. Shareholder preferences and dividend policy. *Journal of Finance* 45:993–1019.

Brennan, M., V. Maksimovic, and J. Zechner. 1988. Vendor financing. *Journal of Finance* 43:1127–1141.

Burkart, M. and T. Ellingsen. 2004. In-kind finance: a theory of trade credit. *American Economic Review* 94:569–590.

Burkart, M., D. Gromb, and F. Panunzi. 1996. Debt design, liquidation value, and monitoring. Mimeo, MIT.

——. 1997. Large shareholders, monitoring and the value of the firm. *Quarterly Journal of Economics* 112:693–728.

Burkart, M., F. Panunzi, and A. Shleifer. 2003. Family firms. *Journal of Finance* 58:2167–2202.

Calomiris, C., C. Himmelberg, and P. Wachtel. 1995. Commercial paper and corporate finance: a microeconomic perspective. *Carnegie-Rochester Series on Public Policy* 42: 203–250.

Carey, M., S. Prowse, J. Rea, and G. Udell. 1993. Recent developments in the market for privately placed debt. *Federal Reserve Bulletin* February:77–92.

Chemla, G., M. Habib, and A. Ljungqvist. 2004. An analysis of shareholder agreements. Mimeo, Imperial College, London, University of Zurich, and New York University.

Chemmanur, T. J. and P. Fulghieri. 1999. A theory of the going-public decision. *Review of Financial Studies* 12:249–279.

Chen, H. C. and J. Ritter. 2000. The seven percent solution. *Journal of Finance* 55:1105–1132.

Choe, H., R. Masulis, and V. Nanda. 1993. Common stock offerings across the business cycle: theory and evidence. *Journal of Empirical Finance* 1:3–31.

Corbett, J. and T. Jenkinson. 1994. The financing of industry, 1970–1989: an international comparison. CEPR DP 948.

Dewatripont, M. and J. Tirole. 1994a. *The Prudential Regulation of Banks*. Cambridge, MA: MIT Press.

——. 1994b. A theory of debt and equity: diversity of securities and manager–shareholder congruence. *Quarterly Journal of Economics* 109:1027–1054.

Dow, J., G. Gorton, and A. Krishnamurthy. 2003. Equilibrium asset prices under imperfect corporate control. National Bureau of Economic Research, Working Paper 9758.

Eckbo, B. E. 1986. Valuation effects of corporate debt offerings. *Journal of Financial Economics* 15:119–151.

Eckbo, E. and R. Masulis. 1992. Cost of equity issuance. In *The New Palgrave Dictionary of Money and Finance* (ed. P. Newman, M. Milgate, and J. Eatwell), Volume 1, pp. 496–499. London: Macmillan.

——. 1995. Seasoned equity offerings: a survey. In *Handbook in Operations Research and Management Science: Finance* (ed. R. Jarrow, V. Maksimovic, and B. Ziemba), Volume 9. Amsterdam: North-Holland.

Emerick, D. and W. White. 1992. The case for private placements: how sophisticated investors add value to corporate debt issuers. *Journal of Applied Corporate Finance* 5(3):83–91.

Fama, E. and K. French. 2001. Disappearing dividends: changing firm characteristics or lower propensity to pay? *Journal of Financial Economics* 60:3–43.

Fazzari, S., R. G. Hubbard, and B. C. Petersen. 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 1:141–195.

Finnerty, J. 1993. An overview of corporate securities innovation. In *The New Corporate Finance: Where Theory Meets Practice* (ed. D. Chew). New York: McGraw-Hill.

Frank, M. Z. and V. K. Goyal. 2003. Testing the pecking order of capital structure. *Journal of Financial Economics* 67: 217–248.

——. 2004. Capital structure decisions: which factors are reliably important? (February 11, 2004). EFA 2004 Maastricht Meetings Paper 2464; Tuck Contemporary Corporate Finance Issues III Conference Paper.

Gertler, M. and S. Gilchrist. 1993. The role of credit market imperfections in the monetary transmission mechanism. *Scandinavian Journal of Economics* 95:43–64.

——. 1994. Monetary policy, business cycle and the behavior of small business firms. *Quarterly Journal of Economics* 109:309–340.

Gibson, S., J. Kose, and L. Lang. 1990. Troubled debt restructuring: an empirical study of private reorganization of firms in default. *Journal of Financial Economics* 27:315–353.

Gompers, P. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50:1461–1489.

——. 1998. An examination of convertible securities in venture capital investments. Harvard Business School, Working Paper.

Gompers, P. and J. Lerner. 1999. *The Venture Capital Cycle*. Cambridge, MA: MIT Press.

——. 2001. *The Money of Invention: How Venture Capital Creates New Wealth*. Boston, MA: Harvard Business School Press.

——. 2003. The really long-run performance of initial public offerings: the pre-Nasdaq evidence. *Journal of Finance* 58:1355–1392.

Graham, J. and C. Harvey. 2001. The theory and practice of corporate finance: evidence from the field. *Journal of Financial Economics* 60:187–243.

Greenbaum, S. and A. Thakor. 1995. *Contemporary Financial Intermediation*. Fort Worth, TX: Dryden Press, Harcourt Brace College Publishers.

Grullon, G. and D. Ikenberry. 2000. What do we know about share repurchases? *Journal of Applied Corporate Finance* 13(1):31–51.

Hanley, K. and J. Ritter. 1992. Going public. In *The New Palgrave Dictionary of Money and Finance* (ed. P. Newman, M. Milgate, and J. Eatwell), Volume 2, pp. 248–255. London: Macmillan.

Harris, M. and A. Raviv. 1988. Corporate control contests and capital structure. *Journal of Financial Economics* 20: 55–88.

Hart, O. and J. Moore. 1989. Default and renegotiation: a dynamic model of debt. Mimeo, MIT and LSE. (Published in *Quarterly Journal of Economics* (1998) 113:1–42.)

Harvard Business School. 1987. Note on financial contracting: deals. Case 9-288-014, rev. 1989.

——. 1990. Note on acquiring bank credit. Case 9-391-010, prepared by P. Bilden.

——. 1991. Note on bank loans. Case 9-291-026, prepared by S. Roth, rev. 1993.

Hoshi, T., A. Kashyap, and D. Scharfstein. 1990. The role of banks in reducing the costs of financial distress in Japan. *Journal of Financial Economics* 27:67–88.

——. 1991. Corporate structure, liquidity and investment: evidence from Japanese industrial groups. *Quarterly Journal of Economics* 106:33–60.

Hubbard, R. 1998. Capital-market imperfections and investment. *Journal of Economic Literature* 36:193–225.

Ibbotson, R. 1975. Price performance of common stock new issues. *Journal of Financial Economics* 2:235–272.

Jagannathan, M., C. P. Stephens, and M. S. Weisbach. 2000. Financial flexibility and the choice between dividends and stock repurchases. *Journal of Financial Economics* 57: 355–384.

James, C. 1987. Some evidence on the uniqueness of bank loans. *Journal of Financial Economics* 19:217–235.

James, C. and P. Weir. 1991. Borrowing relationships, intermediation, and the cost of issuing public securities. *Journal of Financial Economics* 28:149–172.

Jensen, M. and W. Meckling. 1976. Theory of the firm: managerial behavior, agency costs, and capital structure. *Journal of Financial Economics* 3:305–360.

Jolls, C. 1998. Stock repurchases and incentive compensation. National Bureau of Economic Research, Working Paper 6467.

Joskow, P., N. Rose, and A. Shepard. 1993. Regulatory constraints on CEO compensation. *Brookings Papers on Economic Activity, Microeconomics*, pp. 1–58. Brookings Institution Press.

Kahan, M. and B. Tuckman. 1993. Private vs public lending: evidence from covenants. Mimeo, New York University.

Kang, J. K. and R. Stulz. 1994. How different is Japanese corporate finance? An investigation of the information content of new securities issues. National Bureau of Economic Research, Working Paper 4908.

Kaplan, S. and P. Strömberg. 2003. Financial contracting theory meets the real world: an empirical analysis of venture capital contracts. *Review of Economic Studies* 70:281–315.

Kaplan, S. and P. Strömberg. 2004. Characteristics, contracts, and actions: evidence from venture capitalist analyses. *Journal of Finance* 59:2177–2210.

Kaplan, S. N. and L. Zingales. 1997. Do investment-cash flow sensitivities provide useful measures of financing constraints? *Quarterly Journal of Economics* 112:169–216.

——. 2000. Investment-cash flow sensitivities are not valid measures of financing constraints. *Quarterly Journal of Economics* 115:707–712.

Kashyap, A. and J. Stein. 2000. What do a million observations on banks say about the transmission of monetary policy? *American Economic Review* 90:407–428.

Kashyap, A., J. Stein, and D. Wilcox. 1993. Monetary policy and credit conditions: evidence from the composition of external finance. *American Economic Review* 83:78–98.

Kojima, K. 1994. An international perspective on Japanese corporate finance. RIEB DP45, Kobe University.

Kotaro, T. 1995. *The Japanese Market Economy System: Its Strengths and Weaknesses.* Tokyo: LTCB International Library Foundation.

Krigman, L., W. Shaw, and K. Womack. 2001. Why do firms switch underwriters? *Journal of Financial Economics* 60: 245–284.

Lamont, O. 1997. Cash flow and investment: evidence from internal capital markets. *Journal of Finance* 52:83–109.

Lerner, J. 2000. *Venture Capital and Private Equity: A Casebook.* New York: John Wiley.

Lerner, J. and U. Malmendier. 2004. Contractibility and the design of research agreements. Mimeo, Harvard University and Stanford University.

Lerner, J. and R. Merges. 1998. The control of technology alliances: an empirical analysis of the biotechnology industry. *Journal of Industrial Economics* 46:125–156.

Light, J. and W. White. 1979. *The Financial System.* Homewood, IL: Irwin.

Lintner, J. 1956. Distribution of incomes of corporations among dividends, retained earnings, and taxes. *American Economic Review* 46:97–113.

Loughran, T. and J. Ritter. 2002. Why don't issuers get upset about leaving money on the table in IPOs? *Review of Financial Studies* 15:413–444.

Lummer, S. L. and J. J. McConnell. 1989. Further evidence on the bank lending process and the reaction of the capital-market to bank loan agreements. *Journal of Financial Economics* 25:99–122.

Masulis, R. 1988. *The Debt/Equity Choice.* Cambridge, MA: Ballinger Publishing Company.

Masulis, R. and A. Korwar. 1986. Seasoned equity offerings. An empirical investigation. *Journal of Financial Economics* 15:91–117.

Mayer, C. 1988. New issues in corporate finance. *European Economic Review* 32:1167–1189.

Mayer, C. 1990. Financial systems, corporate finance, and economic development. In *Asymmetric Information, Corporate Finance, and Investment* (ed. G. Hubbard). National Bureau of Economic Research, University of Chicago Press.

Megginson, W. and K. Weiss. 1991. Venture capitalist certification in initial public offerings. *Journal of Finance* 46: 879-903.

Michaely, R. and K. Womack. 1999. Conflict of interest and the credibility of underwriter analyst recommendations. *Review of Financial Studies* 12:653-686.

Miller, M. and F. Modigliani. 1961. Dividend policy, growth and the valuation of shares. *Journal of Business* 34:411-433.

Modigliani, F. and M. Miller. 1958. The cost of capital, corporate finance, and the theory of investment. *American Economic Review* 48:261-297.

Myers, S. C. 1984. The capital structure puzzle. *Journal of Finance* 39:575-592.

Oliner, S. and G. Rudebusch. 1993. Is there a bank credit channel to monetary policy? Mimeo, Federal Board of Governors.

Olivier, J. 2000. Growth-enhancing bubbles. *International Economic Review* 41:133-151.

Pagano, M. 1993. The flotation of companies on the stock market: a coordination failure model. *European Economic Review* 37:1101-1125.

Pagano, M., F. Panetta, and L. Zingales. 1998. Why do companies go public? An empirical analysis. *Journal of Finance* 53:27-64.

Pastor, L. and P. Veronesi. 2005. Rational IPO waves. *Journal of Finance* 60:1713-1757.

Petersen, M. and R. Rajan. 1997. Trade credit: theory and evidence. *Review of Financial Studies* 10:661-691.

Philippon, T. 2003. Corporate governance over the business cycle. Mimeo, New York University.

Poterba, J. 1988. Coments on Fazzari, Hubbard and Petersen. *Brookings Papers on Economic Activity*, pp. 200-204. Brookings Institution Press.

Rajan, R. and L. Zingales. 1995. What do we know about capital structure? Some evidence from international data. *Journal of Finance* 50:1421-1460.

——. 2003. The great reversals: the politics of financial development in the 20th century. *Journal of Financial Economics* 69:5-50.

Ritter, J. 1987. The cost of going public. *Journal of Financial Economics* 19:269-282.

——. 2003. Investment banking and securities issuance. In *Handbook of the Economics of Finance* (ed. G. Constantinides, M. Harris, and R. Stulz). Amsterdam: North-Holland.

Ritter, J. and I. Welch. 2002. A review of IPO activity, pricing, and allocations. *Journal of Finance* 57:1795-1828.

Rock, K. 1986. Why new issues are underpriced. *Journal of Financial Economics* 15:187-212.

Sahlman, W. 1990. The structure and governance of venture-capital organizations. *Journal of Financial Economics* 27: 473-521.

Shleifer, A. and R. Vishny. 2003. Stock market driven acquisitions. *Journal of Financial Economics* 70:295-311.

Smith, C. and J. Warner. 1979. On financial contracting: an analysis of bond covenants. *Journal of Financial Economics* 7:117-161.

Smith, J. 1987. Trade credit and informational asymmetry. *Journal of Finance* 42:863-872.

Sraer, D. and D. Thesmar. 2004. Performance and behavior of family firms: evidence from the French stock market. Mimeo, CREST, INSEE.

Stein, J. 2003. Agency, information and corporate investment. In *Corporate Finance: Handbook of the Economics of Finance* (ed. G. Constantinides, M. Harris, and R. Stulz), pp. 111-165. Amsterdam: North-Holland.

Stigum, M. 1990. *The Money Market*, 3rd edn. New York: Irwin.

Titman, S. and R. Wessels. 1988. The determinants of capital structure choice. *Journal of Finance* 43:1-19.

Ventura, J. 2004. Economy growth with bubbles. Mimeo, Centre de Recerca en Economia Internacional, Universitat Pompeu Fabra, and CEPR.

White, M. 1989. The corporate bankruptcy decision. *Journal of Economic Perspectives* 3:129-152.

White, L. 1991. *The S&L Debacle: Public Policy Lessons for Bank and Thrift Regulation.* Oxford University Press.

Willis, J. and D. Clark. 1993. An introduction to mezzanine finance and private equity. In *The New Corporate Finance: Where Theory Meets Practice* (ed. D. Chew). New York: McGraw-Hill.

Wilner, B. 1994. The interest rates implicit in trade credit discounts. Mimeo, Kellogg School, Northwestern University.

Yosha, O. 1995. Information, disclosure costs and the choice of financing source. *Journal of Financial Intermediation* 4: 3-20.

Zimmermann, C. 1975. An approach to writing loan agreement covenants. In *Journal of Commercial Bank Lending*, pp. 213-228.

Zingales, L. 1994. The value of the voting right: a study of the Milan Stock Exchange. *Review of Financial Studies* 7: 125-148.

——. 1995. Inside ownership and the decision to go public. *Review of Economic Studies* 62:425-448.

4 TVEGANJE IN DONOSNOST

4-1 Donos in donosnost

Cilj investitorja je doseči kar se da velik donos na vložena sredstva. Donos investicije je sestavljen iz dveh delov, donosa iz naslova obresti (dividend) in donosa iz naslova kapitalskega dobička oziroma izgube. Pazljivi moramo biti pri uporabi besed donos in donosnost oziroma stopnja donosa. Donos označuje število denarnih enot, ki jih prejme investitor v določenem obdobju. Donosnost pa je donos, izražen relativno (torej v %) glede na investiran znesek.

---

donos: uspeh kake pridobitvene dejavnosti v določenem obdobju, izražen v denarju

donosnost (stopnja donosa): razmerje med donosom in vloženim kapitalom

Slovar slovenskega knjižnega jezika, DZS 1994, str. 159

Različne oblike naložb prinašajo različne donose. V spodnji tabeli vidimo, da je investitor, ki je leta 1925 investiral 1000 dolarjev v delnice velikih ameriških podjetij, imel leta 1995 premoženje vredno 1.114.000 dolarjev.

Tabela 4-1: Vrednost naložbe 1000 dolarjev po 70. letih (1925-95)

Velika podjetja	1.114.000
Manjša podjetja	3.822.000
Dolgoročne državne obveznice	34.000
Kratkoročne obveznice	13.000

Vir: Brigham, Houston, 1998, str. 155.

Kot vidimo, so bili donosi različnih naložb v tem preteklem obdobju različni. Zakaj bi potemtakem kdo sploh investiral na primer v dolgoročne državne obveznice?

4-2 Tveganje

Poleg donosa je pri investicijskih odločitvah ključna kategorija tveganje. O tveganju govorimo takrat, ko prihodnji denarni donosi niso poznani z gotovostjo temveč samo z verjetnostno porazdelitvijo. Tveganje je torej verjetnost, da bo dejanski donos drugačen od pričakovanega.

Bistveno je razumevanje, da lahko tveganje vpliva na donos investitorja negativno ali pa pozitivno, torej je lahko dejanski donos manjši ali večji od pričakovanega.

Investitorje glede na njihov odnos do tveganja razdelimo v tri skupine

- **ljubitelji tveganja** (*risk lovers*): bolj tvegane naložbe jim prinašajo večjo koristnost, ceteris paribus,
- **tveganju nenaklonjeni investitorji** (*risk averse investors*): manjše tveganje jim prinaša večjo koristnost, ceteris paribus,
- **nevtralni investitorji** (*risk neutral investors*): pri oceni investicij ne upoštevajo tveganja

V povprečju investitorji tveganju niso naklonjeni in imajo najraje naložbe, ki imajo veliko verjetnost, da bodo pričakovani donosi enaki dejanskim. V okviru predmeta Poslovne finance – če ni drugače navedeno - vedno obravnavamo tveganju nenaklonjene investitorje. Za te investitorje velja, da so pripravljeni nositi določeno tveganje le, če so ustrezno nagrajeni, kar pomeni, da je zahtevana stopnja donosa ustrezno visoka in da je posledično cena naložbe ustrezno nizka. Malo tvegane naložbe dosegajo višje cene kot druge, bolj tvegane.

tveganju nenaklonjeni investitorji

$\uparrow$ tveganje $\Rightarrow$ $\uparrow$ zahtevana donosnost in $\downarrow$ cena

Ko se investitorji odločajo med dvema naložbami, se tako odločijo za naložbo, ki ima:

- pri enakem pričakovanem donosu manjše tveganje
- pri enakem tveganju večji pričakovani donos.

Tabela 4-2: Donosnost in tveganje (ZDA, 1926-1995) – v odstotkih

	Povprečna letna donosnost	Standardni odklon
Velike delniške družbe	12,7	20,4
Male delniške družbe	17,7	34,4
Dolgoročne podjetniške obveznice	6	8,7
Dolgoročne državne obveznice	5,5	9,2
Kratkoročne državne obveznice	3,8	3,3
Inflacija	3,2	4,6

Vir: Brigham, Houston, 1998, str. 167.

Analiza tveganja je tako sestavni del investicijske odločitve. Tveganje naložbe lahko ocenjujemo z dveh vidikov:

- samostojno tveganje posamezne naložbe (*stand alone risk*)
- tveganje naložbe kot dela premoženja

4-3 Merjenje tveganja posamezne naložbe

Verjetnostna porazdelitev: je seznam možnih izidov in njihovih verjetnosti, da se zgodijo. V tabeli je prikazana verjetnostna porazdelitev možnih stopenj donosa za podjetje A.

Tabela 4-3: Verjetnostna porazdelitev možnih stopenj donosa za podjetje A

Verjetnost	**Stopnja donosa**
p_i	r_A
0,30	100%
0,40	15%
0,30	-70%
vsota=1,00	

Pričakovana stopnja donosa: tista stopnja donosa, ki je najbolj verjetna oziroma tehtano povprečje možnih stopenj donosa

$$E(r) = \Sigma\ p_i * r_i$$

kjer je p_i verjetnost nastopa možne stopnje donosnosti r_i. Vsota verjetnosti p_i je vedno 1. E(r) je torej matematično upanje slučajne spremenljivke r.

Tabela 4-4: Pričakovana stopnje donosa oziroma matematično upanje donosnosti

p_i	r_i	$p_i * r_i$
0,30	1	0,30
0,40	0,15	0,06
0,30	-0,70	-0,21
vsota=1,00		***E(r) = 0,15***

Varianca in standardni odklon: merita tveganja, ki kažeta, koliko dejanske (možne) stopnje donosna odstopajo od pričakovane stopnje donosa. Večje ko je možno odstopanje možnih donosnosti od pričakovane donosnosti, večje je tveganje. Z drugimi besedami, »širša« ko je verjetnostna porazdelitev možnih stopnje donosnosti, bolj tvegana je naložba.

$$\sigma^2 = \sum_{i=1}^{n} (r_i - E(r))^2 \cdot p_i$$

$$\sigma = \sqrt{\sigma^2}$$

Tabela 4-5: Izračun variance

p_i	r_i	$(r_i - E(r))$	$(r_i - E(r))^2$	$(r_i - E(r))^2 * p_i$
0,30	1	0,85	0,72	0,217
0,40	0,15	0	0	0
0,30	-0,70	-0,85	0,26	0,217
vsota=1,00				σ^2 ***=0,43***

$$\sigma = \sqrt{\sigma^2} = \sqrt{0,43} = 0,658 = 65,8\%$$

Uporaba variance oziroma standardnega odklona kot mere tveganja je ustrezna, če se možne stopnje donosov porazdeljujejo kolikor toliko skladno z normalno porazdelitvijo. Tedaj je verjetnost, da se bo dejanska donosnost nahajala v pasu +/- 1σ, 68,26% (+/- 2σ 95,46% in +/- 3σ 99,74%).

Če pričakovani stopnji donosa dveh potencialnih naložb nista enaka, je za primerjavo tveganj teh dveh naložb smiselno uporabiti koeficient variacije.

Koeficient variacije: je delež standardnega odklona v pričakovani stopnji donosa oziroma delež tveganja na enoto donosnosti.

$$KV = \frac{\sigma_i}{E(r_i)};$$ v našem primeru 4,39.

4-4 Merjenje tveganja – tveganje premoženja

Investitorji ponavadi ne naložijo vsega denarja le v eno tvegano naložbo, saj lahko v primeru slabega rezultata vse izgubijo. Z razpršitvijo oziroma diverzifikacijo naložb se zmanjša tveganje celotnega premoženja – pričakovan donos premoženja postane bolj predvidljiva oziroma zmanjša se standardni odklon premoženja.

"Ne nosite vseh jajc v eni sami košari!"

Večina finančnih naložb je del posameznega premoženja. Tako nas ne zanima spreminjanje donosnosti posamezne naložbe, temveč celotnega premoženja. Zato je pomembno, kako posamezna naložba, ki jo vključimo v naše premoženje, vpliva na spreminjanje donosnosti premoženja.

Pričakovana stopnja donosa premoženja: tehtano povprečje pričakovanih donosnosti posameznih komponent (delnic) premoženja

$$E(r) = \Sigma w_i * r_i$$

kjer je w_i utež oziroma delež celotnega premoženja, naložen v posamezno naložbo (delnico).

Tako tveganje premoženja ni povprečje tveganj posameznih naložb, ki ga sestavljajo, temveč je praviloma manjše. Teoretično bi bilo možno sestaviti premoženje tveganih naložb, ki bi bilo samo netvegano. Korelacija med naložbami, ki jih združimo v premoženje, je tisti dejavnik, ki definira tveganje premoženja.

Kovarianca in korelacijski koeficient: varianca in standardni odklon nista odvisni zgolj od varianc in uteži posameznih sestavnih delov premoženja, temveč tudi od njihovih korelacijskih koeficientov oziroma kovarianc. Korelacijski koeficient ($R_{1,2}$), ki lahko zavzame vrednosti med –1 in 1, pove, kako se bodo gibale donosnosti dveh naložb (delnic):

$R_{1,2} > 0$: pozitivna korelacija, donosnost obeh naložb se bo spreminjala v isti smeri
$R_{1,2} < 0$: negativna korelacija, donosnost naložb se bo spreminjala v obratni smeri

$$Cov_{1,2} = \sum_{i=0}^{n} p_i \cdot \left[r_{1,i} - E(r_1) \right] \cdot \left[r_{2,i} - E(r_2) \right]$$

$$R_{1,2} = \frac{Cov_{1,2,}}{\sigma_1 \cdot \sigma_2}$$

Če v premoženje združimo naložbi, ki sta popolno pozitivno korelirani ($R_{1,2}$ = 1), ne zmanjšamo tveganja premoženja. Če pa v premoženje združimo naložbi, ki sta popolno negativno korelirani ($R_{1,2}$ = -1), bo premoženje popolnoma netvegano oziroma uspela nam je popolna razpršitev. Dejansko so pari delnic večinoma pozitivno korelirani ($R_{1,2}$ se giblje med 0,5 in 0,7), kar pomeni, da razpršitev premoženja ne odpravi tveganja, ga pa zmanjša.

Varianca in standardni odklon: merili tveganja, ki kažeta, koliko dejanske (možne) stopnje donosa odstopajo od pričakovane stopnje donosa.

varianca: $\sigma^2 = \sum_{i=1}^{n} (r_i - E(r))^2 \cdot p_i$

standardni odklon: $\sigma = \sqrt{\sigma^2}$

4-5 Primer razpršitve

Kot primer poglejmo, kaj se zgodi, če naše premoženje oblikujemo iz dveh delnic, ki imata enako pričakovano donosnost in enak standardni odklon. Predpostavimo korelacijski koeficient $r_{A,B}$=0,4.

Tabela 4-6: Naložbi, med kateri lahko razpršimo svoje premoženje

	E(r)	σ
Podjetje A	13%	30%
Podjetje B	13%	30%

$$E(r_p) = w_A \cdot E(r_A) + w_B \cdot E(r_B) = 0,5 \cdot 13\% + 0,5 \cdot 13\% = 13\%$$

Pričakovana donosnost premoženja je seveda 13%. Je potem sploh smiselno kombinirati delnici A in B ter iz njiju sestaviti premoženje?

$$\sigma_p = \sqrt{w_A^2 \cdot \sigma_A^2 + w_B^2 \cdot \sigma_B^2 + 2 \cdot w_A \cdot w_B \cdot \sigma_A \cdot \sigma_B \cdot r_{A,B}} = \sqrt{0,063} = 25,1\%$$

Če delnici podjetij A in B nista popolnoma pozitivno korelirani (kar je zelo malo verjetno), potem mora biti njuna »kombinacija« za investitorja ugodnejša odločitev kot naložba samo v eno od obeh delnic. V našem primeru je standardni odklon premoženja za skoraj 5 odstotnih točk manjši od standardnih odklonov delnic A in B.

4-6 Razpršitev, sistematično in nesistematično tveganje

Z naključnim dodajanjem novih naložb (delnic) v premoženje investitorja se hitro zmanjšuje standardni odklon donosnosti premoženja.

Vključitev prvih 5 do 10 naložb v premoženje prinese veliko večje zmanjšanje tveganja kot kasnejše dodajanje naložb. Popolne razpršitve, s katero bi tveganje popolnoma odpravili, ne moremo doseči, saj donosnosti naložb med seboj niso popolnoma negativno korelirani.

skupno tveganje = nesistematično tveganje + sistematično tveganje

Nesistematično tveganje (*unsystematic risk, diversifiable risk, asset-specific risk*): tveganje premoženja, ki ga lahko odpravimo z razpršitvijo.

Sistematično tveganje (*systematic risk, nondiversifiable risk, market risk*): preostalo tveganje, ki je odvisno od splošnih gospodarskih pogojev in ga z razpršitvijo ni mogoče odpraviti.

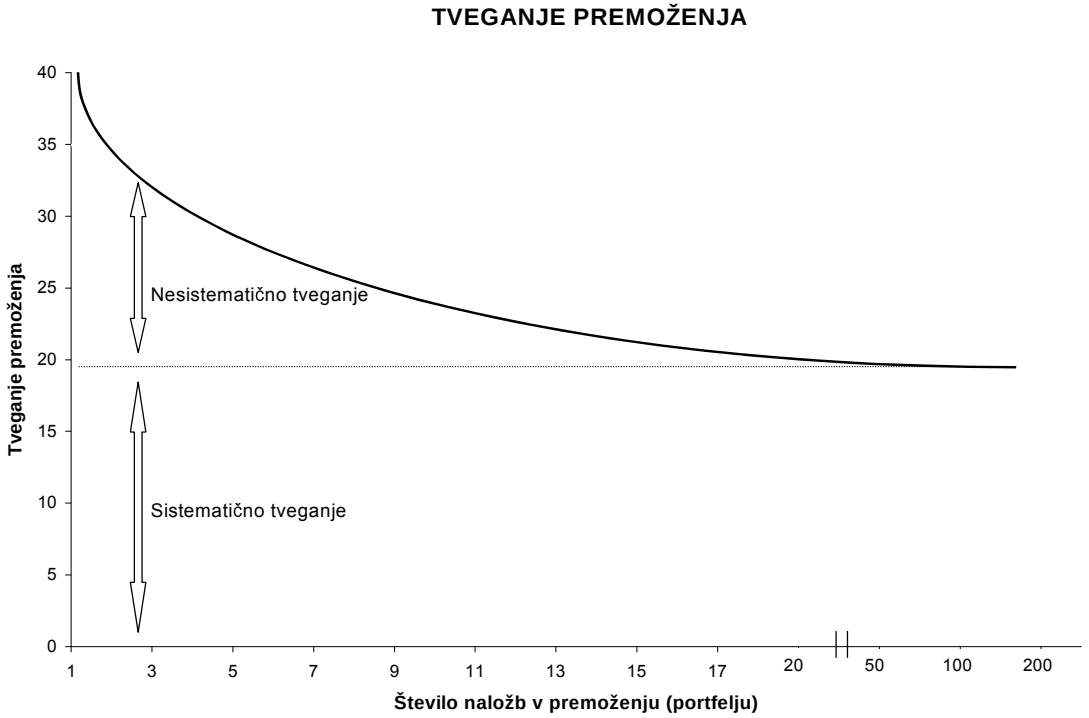
V tabeli Tabela 4-7 so prikazana različna premoženja z naključno izbranimi delnicami z newyorške borze (NYSE).

Tabela 4-7: Število delnic v premoženju in standardni odkloni (v %) – primer NYSE

Število delnic v premoženju	Povprečni standardni odklon donosnosti
1	49,24
2	37,36
4	29,69
8	24,98
20	21,68
50	20,20
100	19,69
200	19,42
400	19,29
1000	19,21

Vir: Ross, Westerfield, Jaffe, 1993, str. 380

Slika 4-1: Tveganje premoženja – sistematično in nesistematično tveganje



Vir: Prirejeno po Ross, Westerfield, Jaffe, 1993, str. 381

4-7 Tržno premoženje

Ob predpostavki racionalnega obnašanja investitorjev so njihova premoženja čimbolj optimalno razpršena in imajo potemtakem skoraj enake značilnosti kot tržno premoženje. Tržno premoženje je pomanjšana slika vseh možnih naložb na kapitalskem trgu. Teoretično investitor razporedi svoj denar na prav vse naložbe, ki obstajajo, in to v takih razmerjih, s kakršnimi so posamezne naložbe udeležene v celotnem tržnem premoženju. Praktično zadostuje že manjša diverzifikacija (na primer prek vzajemnega sklada), ki razprši vloženi denar na nekaj deset naložb.

Od netvegane naložbe, ki jo sestavljajo različni državni vrednostni papirji, investitor zahteva ***netvegano stopnjo donosa*** (r_f). Od tvegane naložbe, ki jo predstavlja tržno premoženje, pa zaradi večjega tveganja zahteva nekaj več, in sicer ***tržno premijo za tveganje*** ($r_m - r_f$). Tržna premija za tveganje je razlika med zahtevanima donosnostima tržnega premoženja in netveganega premoženja ($r_m - r_f$).

Tržna cena tveganja premoženja (λ) je obseg tržne premije za tveganje na enoto tveganja,

$$\lambda = \frac{r_m - r_r}{\sigma_m}$$

Zahtevna stopnja donosa premoženja: $r_p = w_m \cdot r_m + (1 - w_m) \cdot r_f$

ob dejstvu, da je $\sigma_p = w_m \cdot \sigma_m$

sledi: $r_p = r_f + \left(\frac{r_m - r_f}{\sigma_m} \right) \cdot \sigma_p$ oziroma $r_p = r_f + \lambda \cdot \sigma_p$

4-8 Obseg tveganja posamezne naložbe kot dela premoženja

Sprašujemo se, kako meriti tveganje delnice, ki jo vključujemo v naše premoženje. Očitno je relevanten samo tisti del tveganja (standardnega odklona) posamezne naložbe, ki se ga ne da odpraviti z diverzifikacijo. Gre torej za sistematično tveganje posamezne naložbe ($\sigma_{i,s}$):

$$\sigma_{i,s} = \frac{Cov_{i,m}}{\sigma_m}$$

Premijo za tveganje posamezne naložbe zapišemo kot produkt cene tveganja in obsega tveganj posamezne naložbe:

$$pt_i = \frac{r_m - r_r}{\sigma_m} \cdot \frac{Cov_{i,m}}{\sigma_m}$$

$$pt_i = (r_m - r_r) \cdot \frac{Cov_{i,m}}{\sigma_m^2} \ ,$$

kjer drugi člen predstavlja **koeficient beta**: $\beta = \frac{Cov_{i,m}}{\sigma_m^2}$

Koeficient β je zelo uporabna mera tveganja. Meri prispevek posamezne naložbe k tveganju tržnega premoženja. Ali povedano drugače, meri usklajenost gibanja donosnosti delnice z donosnostjo tržnega premoženja.

β tržnega premoženja β_m je tehtana aritmetična sredina bet posameznih naložb in je po definiciji enaka 1. Rečemo lahko tudi, da je β povprečne delnice oziroma naložbe enaka 1.

$$\beta_m = \sum_{i=1}^{n} w_i \cdot \beta_i = 1$$

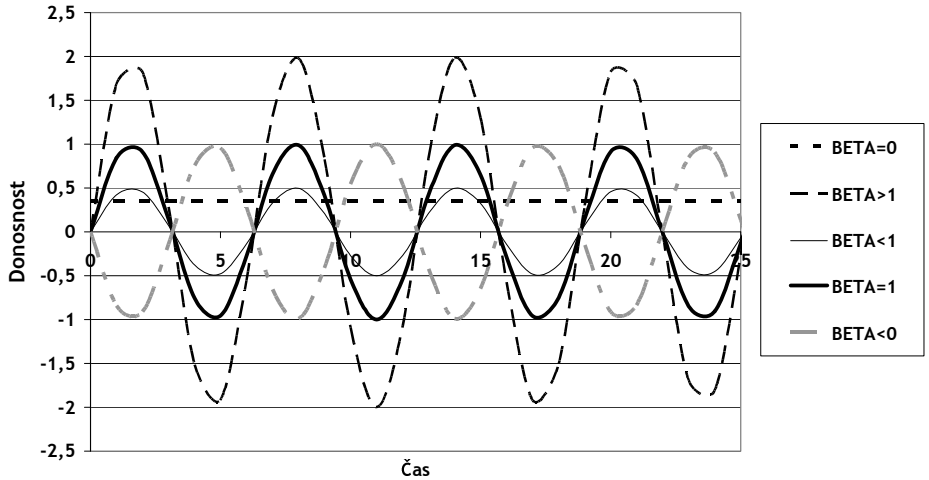
Posamezna naložba z β med 0 in 1 je podpovprečno tvegana. Posamezna naložba z β>1 je nadpovprečno tvegana in bo povečala tveganje premoženja in tudi pričakovano donosnost premoženja, kot je to prikazano na sliki 4-2. Če se donosnost tržnega premoženja poveča za 1%, se bo donosnost te posamezne naložbe povečala za več kot 1%. Teoretično je možna tudi negativna vrednost β; donosnost naložbe bi padala, ko bi donosnosti ostalih naložb (premoženja) naraščale. V praski negativne β še ni bilo. Podatki za ameriški delniški trg kažejo, da se β delniških družb v povprečju gibajmo med 0,75 in 1,50. Vrednost beta koeficienta netvegane naložbe je 0, kar pomeni, da je zahtevana donosnost na takšno naložb enaka netvegani donosnosti r_f.

Tabela 4-8: β nekaterih ameriških podjetij

Delnica	β
America Online	2,10
Microsoft	1,20
General Electric	1,15
Procter&Gamble	1,05
Coca-Cola	1,00
Heinz	0,90
IBM	0,90

Vir: Brigham, Houston, 1998, str. 180.

Slika 4-2: Simoblični prikaz različnih vrednosti koeficienta beta v povezavi z zahtevano donosnostjo naložb skozi čas



4-9 Model določanja cen dolgoročnih naložb (CAPM)

CAPM model (*Capital Asset Pricing Model*) je model za določanje zahtevane donosnosti posamezne naložbe. Model podaja razmerje med tveganjem, ki ga posamezna naložba doprinese k tveganju premoženja (portfelja) in zahtevano donosnostjo te posamezne naložbe.

Enačba, ki kaže razmerje med zahtevano donosnostjo in tveganjem posamezne naložbe (β), je:

$$r_i = r_f + (r_m - r_f) \cdot \beta_i \qquad \beta_i = \frac{Cov_{i,m}}{\sigma_m^2}$$

kjer je $(r_m - r_f)$ tako imenovana **tržna premija za tveganje** posamezne naložbe (*market risk premium*).

Zahtevana donosnost posamezne naložbe je tako sestavljena iz netvegane stopnje donosa in dodatne premije za tveganje, ki ga ta naložba »prinese« v premoženje. Seveda je premija za tveganje nagrada za prevzemanje sistematičnega (tržnega) tveganja, ki se ga ne da odpraviti z razpršitvijo.

CAPM model temelji na "močnih" predpostavkah. Te predpostavke se zelo razlikujejo od stvarnosti. Problemi z modelom so tudi opredelitev tržnega premoženja, prihodnjih možnih in pričakovanih donosnosti tržnega premoženja in netvegane donosnosti. Kljub kritikam CAPM modela je le-ta eden od temeljev določanja cene tveganja.

4-10 Tveganje obveznic

Zahtevana donosnost obveznice je odvisna od različnih tveganj, ki jim je obveznica izpostavljena. Zahtevana donosnost obveznice je sestavljena iz netvegane donosnosti in skupne premije za tveganje, ki jo sestavlja več različnih premij za tveganje.

Tveganje spremembe obrestne mere odraža dejstvo, da se tržna obrestna mera stalno spreminja in da se posledično spreminja tudi zahtevana donosnost. Z naraščanjem ročnosti obveznice narašča tudi *premija za ročnost*.

Pri vseh obveznicah z izjemo državnih oziroma obveznic z državnim jamstvom obstaja možnost, da izdajatelj obveznice ne bo pravočasno in v polnem obsegu izpolnil svojih obveznosti (plačilo obresti in vračilo glavnice):

- *premija za kreditno tveganje*: zahtevana donosnost obveznice in pričakovana povprečna donosnost ob predpostavki neodvisnosti izostankov plačil

- *premija za tržno kreditno tveganje*: predpostavka, da izostanki plačil niso neodvisni od donosnosti tržnega premoženja.

Kot pri vseh naložbah je tudi pri obveznicah pomembna likvidnost. Bolj kot je obveznica nelikvidna, večja je *premija za nelikvidnost*.

Tveganje spremembe obrestne mere in tveganje reinvestiranja

Tveganje spremembe obrestne mere (*interest rate risk*) bi lahko imenovali tudi tveganje spremembe cene. Predstavlja namreč tveganje zmanjšanja cen obveznic, ko obrestne mere naraščajo. Padanje cen je razumljivo, saj investitorji niso več pripravljeni trgovati po nominalnih vrednostih obveznice, če je zahtevana donosnost, ki se pri obveznici odraža preko obrestne mere, previsoka, saj so denarni tokovi, ki jih obveznica prinaša, nespremenljivi. Če investitorji želijo, da jim obstoječi denarni tokovi prinesejo zadovoljivo donosnost, morajo danes za obveznico plačati nižjo ceno. Rečemo torej lahko, da se obveznice, ki imajo nižjo kuponsko obrestno mero, kot znaša tržna obrestna mera, prodajajo z diskontom, torej po ceni, nižji od nominalne vrednosti (*face value*), obveznica, katere kuponska obrestna mera presega tržno obrestno mero, pa s premijo, torej po ceni, ki presega nominalno vrednost.

Če si nadalje ogledamo primer dveh obveznic (Slika 4-3), ki imata enako kuponsko obrestno mero in različni obdobji dospelosti, vidimo, kako sta obveznici različno občutljivi na spremembe tržne obrestne mere. Ugotovimo lahko, da nosijo obveznice z daljšim dospetjem, v primerjavi obveznicam, ki zapadajo hitreje, bistveno višja tveganja obrestne mere. Cena se jim ob enaki spremembi bolj spremeni.

Poleg tveganja obrestne mere pa obveznice nosijo tudi tveganje reinvestiranja (*reinvestment risk*), ki se nanaša na zmožnost vlaganja finančnih prihrankov po vnaprej določeni obrestni meri. Če kupimo obveznico, ki ima do dospetja dvajset let in izdajatelj nima možnosti odpoklica, smo 'gotovi', da se bo vsaj glavnica obrestovala do ob investiranju določeni obrestni meri. Za kuponske obresti to seveda ne velja, saj jih, ko le-te zapadajo, nadalje lahko vlagamo le po tržni obrestni meri. Če pa imamo investicijski horizont krajši, npr. dve leti, pa ob padcu obrestnih mer, tudi glavnico (in ne zgolj obresti) vložimo le po nižjih tržnih donosnostih. Vidimo torej, da nosijo obveznice s krajšim rokom dospetja več tveganja reinvestiranja kot tiste z daljšim rokom dospetja.

Vprašanja in naloge

V 4-1 Koeficient variacije, izračunan kot količnik standardnega odklona in pričakovane donosnosti, je standardizirana mera tveganja na enoto pričakovane donosnosti.

a. Da b. Ne

V 4-2 CAPM model predpostavlja, da investitorje v prvi vrsti zanima tveganje premoženja in ne tveganje posamične naložbe. Zato jih pri posamezni naložbi zanima le doprinos te naložbe k tveganju celotnega premoženja.

a. Da b. Ne

V 4-3 Diverzifikacija premoženja zmanjšuje variabilnost donosov vsake naložbe, ki sestavlja premoženje.

a. Da b. Ne

V 4-4 Kombiniranje dveh naložb v premoženje znižuje tveganje, če le ni popolne korelacije med obema naložbama ($r = -1$; $r = 1$).

a. Da b. Ne

V 4-5 Racionalni investitor naloži denar v naložbo le, če je pričakovana donosnost dovolj visoka, da mu kompenzira (poplača) tveganje naložbe. Investitorji, ki imajo premoženje razpršeno na več naložb, zahtevajo ustrezno premijo oziroma poplačilo le za tisti del tveganja, ki ga ni mogoče odpraviti z razpršitvijo, torej za nesistematični del tveganja.

a. Da b. Ne

V 4-6 Za delnico podjetja Fricko d.d. poznamo naslednjo verjetnostno porazdelitev:

Verjetnost	Donosnost (v %)
0,15	-20
0,20	0
0,30	10
0,20	15
0,15	20

a. Izračunajte pričakovano stopnjo donosa.
b. Izračunajte standardni odklon donosnosti delnice.

V 4-7 Za delnico podjetja Efcom d.d. je poznana naslednja verjetnostna porazdelitev donosnosti:

Verjetnost	Donosnost (v %)
0,05	-10
0,25	-2
0,40	4
0,25	8
0,05	35

a. Koliko znaša pričakovana donosnost delnice podjetja Efcom d.d.?
b. Koliko znaša standardni odklon donosnosti te delnice?
c. Ali se vam zdi delnica tvegana naložba?

V 4-8 Dani sta verjetnostni porazdelitvi donosnosti delnic podjetij Enka d.d. in Dvojka d.d.:

Verjetnost	Donosnost Enka d.d.	Donosnost Dvojka d.d.
0,05	-15%	5%
0,10	4%	15%
0,55	15%	35%
0,20	25%	50%
0,10	45%	100%

a. Izračunajte pričakovano donosnost naložb v delnico podjetij Enka in Dvojka.
b. Katera naložba je bolj tvegana?

V 4-9 Kot študenta financ so vas doma pooblastili, da pametno vlagate družinski denar na borzi vrednostnih papirjev. Trenutno razmišljate o naložbi, za katero so znani naslednji podatki za preteklih 5 let:

Leto	Letna donosnost
1996	10%
1997	9%
1998	7%
1999	6%
2000	6%

Kako boste ocenili naložbo, če veste, da je imela po donosnostih zelo podobna naložba v enakem preteklem obdobju standardni odklon 0,033?

V 4-10 Vaš sosed Filip namerava investirati prihranke v delnice podjetja Muki d.d. in je pridobil naslednje podatke:

Verjetnost	Donosnost (v %)
0,10	-10
0,20	5
0,30	10
0,40	25

a. Filip je izračunal, da je pričakovana donosnost delnice 7,5%. Prosil vas je za preverbo izračuna.
b. Filip vas je tudi prosil, da mu izračunate standardni odklon naložbe.
c. Kakšen sklep lahko skupaj postavita na podlagi izračunanih podatkov?

V 4-11 Po predavanju Poslovnih financ ste skočili še v knjižnico, ker ste naleteli na brošuro s podatki o podjetju P-fin d.d.. Poskušate ugotoviti, kaj lahko na podlagi podatkov iz brošure rečete o tem podjetju.

Leto	Donosnost P-fin d.d.	Tržna donosnost
1995	8,00%	7,80%
1996	8,30%	8,15%
1997	7,50%	7,80%
1998	7,30%	7,70%
1999	7,60%	7,70%
2000	8,30%	8,00%

V 4-12 Donosnost netvegane naložbe je 5%, donosnost povprečno tvegane naložbe na trgu 11%. Beta koeficient delnice podjetja Xport d.d. je 1,3.

a. Kolikšna je zahtevana donosnost naložbe v delnico Xporta?
b. Kako bi se odgovor spremenil, če bi investitorji pričakovali porast inflacije za 2% točki? (ni potrebo upoštevati Fisherjeve enačbe)
c. Kako bi na zahtevano donosnost delnice Xporta vplivalo povečanje tržne premije za tveganje za 3% točke?
d. Kakšen bi bil vpliv obeh sprememb na zahtevano donosnost?

V 4-13 Premoženje imate naloženo v delnicah ugledanih slovenskih podjetij. Netvegana stopnja donosa je 7%. Lani ste dosegli 13% donosnost na premoženje. Zanima vsa zahtevana donosnost premoženja za letos, če se glede na lani nič spremenilo, razen premije za tveganje, ki se je povečala za dve odstotni točki.

Delnica	Vrednost (v EUR)	β
Krkec d.d.	100.000	0,8
Merki d.d.	150.000	1,2
Petko d.d.	50.000	1,8

V 4-14 Investicijski sklad Mladi lev je sestavljen iz 5 delnic, za katere so znani sledeči podatki:

Delnica	Vrednost (v mio EUR)	β
1	130	0,4
2	110	1,5
3	70	3,0
4	90	2,0
5	50	1,0

a. Predpostavite 12% donosnost za netvegane naložbe in 6% tržno premijo za tveganje ter izračunajte β sklada.
b. Kolikšna je zahtevana donosnost sklada?

V 4-15 Kot pripravnik v SloCap finančni hiši upravljate s premoženjem 10 vrednostno enakomerno zastopanih delnic. β premoženja je 1,64, β delnic A, ki je vključena v premoženje, pa je 2. Odločite se prodati vse delnice A in jih nadomestiti z delnicami B. Po zamenjavi je β premoženja 1,55. Koliko je β delnice B?

V 4-16 Kot diplomant Ekonomske fakultete - smer finance ste se zaposlili v Fidelity Investments, ki je nedavno odprl svojo pisarno v Ljubljani. Zaradi tega je njegov investicijski sklad Urko 1 še majhen, obsega vsega 5 naložb, za katere so poznani naslednji podatki:

Naložba	Vrednost (mio EUR)	β
1	115	1,4
2	60	2,5
3	170	1,2
4	400	1,0
5	55	2,0

Vaša prva naloga je bila izračunati zahtevano donosnost sklada. Žal ste šefa razočarali, saj izračuna niste naredili, ker niste imeli podatka o povprečni tržni donosnosti. Naslednji dan ste poskusili znova, pomagali pa ste si z oceno verjetnostne porazdelitve tržne donosnosti, ki ste jo našli v poslovnem dnevniku Finance:

Verjetnost	Tržna donosnost
0,05	0,08
0,10	0,09
0,20	0,11
0,40	0,13
0,15	0,14
0,10	0,16

a. Zapišite ocenjeno enačbo CAPM modela (netvegana stopnja donosa je 3%) in izračunajte zahtevano donosnost sklada.
b. Koliko bi znašala zahtevana stopnja donosa sklada za naslednje obdobje, če bi struktura naložb ostala nespremenjena?
c. Ker ste na zgornji vprašanji pravilno odgovorili, vam je šef naroči, da ocenite naložbo v delnico podjetja Župa d.d.. Pričakovana donosnost te naložbe je 25%, β je 2,8, sklad pa razmišlja o nakupu 40 milijonov tolarjev te delnice.

V 4-17 Podedovali ste premoženje vaše stare mame, ki je sestavljeno iz 12 naložb, od katerih vsaka predstavlja različni del premoženja. Skupna vrednost premoženja je 3,5 milijona tolarjev, β pa znaša 1,15. Odločili ste se preizkusiti svoje finančno znanje in spremeniti strukturo premoženja. Tako boste prodali 5 lotov delnice A, kar predstavlja 1/10 premoženja. β delnice A je 0,95. Z izkupičkom od prodaje nameravate kupiti neko delnico B, ki ima β 1,05. Kako bo ta transakcija vplivala na sistematično tveganje vašega premoženja?

V 4-18 Za seminarsko nalogo pri nekem finančnem predmetu morate pojasniti učinke razpršitve naložb na tveganje in donosnost premoženja. Zbrali se podatke o dveh podjetjih. Za vse možna stanja gospodarstva ste vzeli enako verjetnost.

Stanje gospodarstva	Donosnost Venera d.d.	Donosnost Mars d.d.
depresija	-30%	5%
recesija	15%	15%
prosperiteta	40%	-10%
blaginja	60%	15%

Za pripravo seminarske naloge želite izračunati:

a. pričakovano donosnost naložb v obe delnici.
b. standardna odklona in koeficienta variacije, ki merita tveganje posamezne naložbe.

c. kovarianco in korelacijski koeficient, ki merita povezanost donosnosti obeh naložb.
d. pričakovano donosnost premoženja, ki je sestavljeno iz delnic podjetja Venera (60%) in delnic podjetja Mars (40%).
e. varianco in standardni odklon premoženja.

V 4-19 Po naročilu šefa morate kupiti eno od dveh naložb: delnico podjetja Jabolko d.d. ali delnico podjetja Krompir d.d.. Tržna cena delnice Jabolko je 52,00 EUR. V primeru recesije gospodarstva (verjetnost 50%) bo imela v prihodnjem letu tržno vrednost 59,00 EUR, v primeru stagnacije 70,00 EUR (verjetnost 30%) in v primeru vzpona 95,00 EUR (verjetnost 20%). Znani so še naslednji podatki:

- standardni odklon donosnosti tržnega premoženja je 0,1
- standardni odklon donosnosti delnice Krompir je 0,12
- pričakovana donosnost delnice Krompir je 0,09
- korelacijski koeficient tržne donosnosti in donosnosti delnice Jabolko je 0,8
- korelacijski koeficient tržne donosnosti in donosnosti delnice Krompir je 0,2
- korelacijski koeficient med donosnostjo delnic Jabolko in Krompir je 0,6.

a. Za katero od delnic bi se odločil tveganju nenaklonjen investitor?
b. Kolikšna je pričakovana donosnost premoženja, sestavljenega iz 70% delnic podjetja Jabolko in 30% delnic podjetja Krompir?
c. Kolikšen je koeficient β premoženja?

V 4-20 Ste upravljavec portfelja, sestavljenega iz delnic logističnega (L d.d.) in prehrambenega (P d.d.) podjetja. Ker vas skrbi poslovanje logističnega podjetja, katerega β zanaša 1,34, poleg tega pa se vam dozdeva, da je management vpleten v nastajajočo finančno afero, ste odločeni prodati njegove delnice. Seveda bi zaradi slabe razpoložljivosti naložb vložili v delnico podjetja P d.d.. β premoženja po zamenjavi naložb bi znašala 1,77. Kako je sestavljen vaš portfelj pred prodajo, če veste, da njegova donosnost znaša 16,3%, netvegana stopnja donosa 2,5%, donosnost tržnega premoženja pa 12%?

Odgovori in rešitve nalog

O 4-1 a.

O 4-2 a.

O 4-3 b.

O 4-4 b.

O 4-5 b.

O 4-6 E(r)=6%, σ=12,61

O 4-7

a. 4,35%
b. 8,41%
c. Odvisno od posameznikove nagnjenosti k tveganju.

O 4-8

a. 17,4% in 41%
b. b) KV_E=73%, KV_D=55% $\rightarrow$ Enka d.d. je bolj tvegana

O 4-9 σ=1,62 kar je manj od σ primerljive naložbe (3,3%), torej je analizirana naložba zanimiva.

O 4-10

a. E(r)=13%
b. σ=11,23

O 4-11 Delnica podjetja P-fin d.d. je nadpovprečno tvegana; E(r-P-fin)=7,83%, E(r-tržna)=7,86%, σ(P-fin)=0,39%, σ(tržna)=0,16%, β(P-fin)=2,2

O 4-12

a. r=12,8%
b. r=14,8%
c. r=16,7%
d. r=18,7%

O 4-13 r=15,33%

O 4-14

a. β=1,46
b. r=20,7%

O 4-15 beta nove delnice je 1,1

O 4-16

b. 15,03%
c. Ne, naložba ni zanimiva.

O 4-17 β=1,16

O 4-18

a. r_v=21,25%; r_m=6,25%
b. KV_v=1,58; KV_m=1,64
c. Cov=0,00109; $R_{v,m}$=0,0318
d. r=15,25%
e. σ=20,7%

O 4-19

a. Jabolko
b. 26,17%
c. β=1,528

O 4-20 74,4% delnic L d.d. in 25,6% delnic P d.d.

5 STROŠKI KAPITALA

5-1 Stroški kapitala

V financah s pojmom kapital označujemo vse dolgoročne vire financiranja podjetja: navadni lastniški kapital, prednostni lastniški kapital, dolgoročni dolg (in finančni zakup).

Med najpomembnejše poslovne odločitve v podjetju spadajo odločitve o novih dolgoročnih naložbah (*capital budgeting*). Med tehtanjem o sprejetju oziroma zavrnitvi nove dolgoročne naložbe je ključen podatek o stroških kapitala, ki ga mora poznati oziroma izračunati finančni direktor. Pod pojmom stroški kapitala razumemo strošek oziroma ceno dolgoročnih in trajnih virov financiranja podjetja. Strošek kapitala je le delno pod nadzorom oziroma vplivom podjetja, saj so nekatere vhodne spremenljivke dane od zunaj (na primer obrestne mere in davčne stopnje).

Podjetje zanima samo strošek dodatnega kapitala, ki ga potrebuje za realizacijo novih dolgoročnih naložb in ne strošek že »obstoječega« kapitala.

Koncept stroškov kapitala bomo uporabili pri analizi investicijskih odločitev, sicer pa se uporablja tudi na drugih področjih. V okviru analize investicijskih odločitev temelji koncept stroškov kapitala na dveh predpostavkah:

1. Struktura dolgoročnih finančnih virov bo ostala nespremenjena tudi po odločitvi o novi dolgoročni naložbi. Z drugimi besedami, podjetje bo še naprej imelo tako imenovano ciljno strukturo kapitala
2. Tveganje nove dolgoročne naložbe je enako tveganju obstoječega poslovanja podjetja.

Podjetje tako vedno uporablja tako strukturo financiranja nove naložbe kot jo že sicer ima. Če je na primer strošek financiranja z dolgom 10%, strošek financiranja z lastniškim kapitalom pa 14%, potem je napačno reči, da bo podjetje financiralo novo dolgoročno naložbo le z dolžniškim kapitalom, ki je cenejši. S tem bi namreč podjetje izkoristilo svojo kapaciteto zadolževanja in bi v prihodnje slej ko prej moralo poseči tudi po novem lastniškem kapitalu ter tedaj sprejeti tudi njegovo višjo ceno. Zato se stroške kapitala, uporabljen za investicije, računa kot tehtano povprečje različnih virov financiranja, ki jih uporablja podjetje, ne glede na dejanski način financiranja konkretne dolgoročne naložbe. Od tod izvira izraz za stroške kapitala *tehtano povprečje stroškov kapitala* oziroma WACC (*Weighted Average Cost of Capital*).

Tabela 5-1: WACC 8 ameriških podjetij (1996)

Podjetje	**WACC** (v %)
Intel	14,5
General Electric	13,5
Walt Disney	12,3
Coca-Cola	12,0
Motorola	11,6
AT&T	9,9
Exxon	9,4
Wal-Mart	9,4

Vir: Brigham, Houston, 1998, str. 365.

5-2 Ocenjevanje stroškov kapitala

Podjetja se načeloma financirajo z različnimi vrstami kapitala. Ne glede na vrsto je kapital produkcijski faktor in zato nekaj stane (ni zastonj). Prvi korak pri izračunu stroškov kapitala je tako določitev stroškov posameznih vrst kapitala.

Kapital ločimo na dolžniški in lastniški kapital, pri čemer lastniški kapital delimo na prednostne delnice in navadni lastniški kapital, ki ga sestavljajo osnovni kapital, zadržani dobički in vplačan presežek kapitala. Da bi lahko prišli do ocene WACC, moramo torej analizirati in določiti strošek vsake od vrst kapitala. Te stroške potem pomnožimo z deleži (utežmi) posamezne vrste kapitala v celotnem kapitalu podjetja in dobimo WACC:

$$WACC = w_d\ r_d\ (1\text{-}T)\ +\ w_{ps}\ r_{ps} + w_s\ r_s$$

V nadaljevanju bomo pogledali načine ocenjevanja stroškov dolga, prednostnih in navadnih delnic.

5-3 Strošek dolga

Podjetje pride do dolžniškega kapitala z najetjem kredita (načeloma pri banki, lahko pa tudi drugje), zakupom ali pa z izdajo obveznic oziroma drugih dolžniških papirjev. Pri najetju kredita je strošek kapitala obrestna mera, ki jo zaračuna banka, k čemur moramo prišteti še druge stroške (zavarovanje kredita, stroški obdelave zahteve in podobno).

Če ima podjetje že izdane obveznice, potem lahko s pomočjo podatkov o teh »starih« obveznicah ugotovi ceno, po kateri se lahko zadolžuje. S pomočjo podatkov o nominalni vrednosti obveznic, kuponski obrestni meri oziroma kuponu, časom do dospetja in trenutni ceni, lahko finančni direktor podjetja izračuna zahtevano donosnost.

Toda sama zahtevana donosnost obveznic še ne predstavlja celotnega stroška dolga podjetja. Potrebno je upoštevati še stroške izdaje nove serije obveznic (na primer strošek investicijske banke, ki pripravi in izvede samo izdajo, stroški različnih dovoljenj, stroški pravnikov in drugo)[11]. Podjetje mora te stroške odšteti od izkupička od prodaje novih obveznic in posledično so stroški dolžniškega kapitala ustrezno višji.

Pri financiranju z dolgom moramo upoštevati tudi davke, ki imajo z vidika podjetja »ugoden« učinek na strošek dolga. Obresti na dolg so namreč odbitna postavka od davčne osnove in znižujejo davčno breme podjetja:

stroški dolga po davkih =obresti - davčni prihranki

oziroma

*stroški dolga po davkih = stroški dolga pred davki * (1 - davčna stopnja)*

Vzemimo, da je strošek dolžniškega kapitala za podjetje 10% in da ima podjetje dejansko davčno stopnjo 25%. Strošek dolga po davkih je občutno nižji (7,5%):

$$r_{d,at} = r_d * (1-T)$$

$$r_{d,at} = 10\% * (1-0,25) = 7,5\%$$

Za podjetje so torej relevantni stroški dolga po davkih ($r_{d,at}$, kjer indeks *at* pomeni *after tax*), kar je za podjetje ugodno, saj so ti stroški nižji od stroškov dolga pred davki.

5-4 Stroški prednostnih delnic

Prednostnih delnic (*preferred stocks*) je več vrst. Tako na primer prednostne delnice s fiksno in kumulativno dividendo izplačajo imetnikom vsako leto enako dividendo. Če poslovni rezultat podjetja izplačila dividend ne dovoljuje, se neizplačane dividende kumulirajo in se v celoti izplačajo s tekočo dividendo v letu, ko to poslovni rezultat ponovno omogoča. Z vidika tveganja so prednostne delnice bolj tveganje kot obveznice (dolg) podjetja. Ob predpostavki tveganju nenaklonjenih investitorjev, ki nas spremlja skozi ves tekst, to za podjetje pomeni večji strošek prednostnih delnic v primerjavi z obveznicami.

Tudi z davčnega vidika podjetja so prednostne delnice manj ugodne, saj dividende niso odbitna postavka od osnove za izračun davka od dobička podjetja.

Ob predpostavki, da ima podjetja že izdane prednostne delnice, za katere poznamo trenutno tržno ceno (P_0) in dividendo (D), lahko izračunamo strošek kapitala, ki bi ga zbrali z izdajo nove serije prednostnih delnic.

[11] Podrobneje stroške izdaje obveznic obravnava poglavje o financiranju z dolgoročnim dolgom.

Seveda moramo tako kot pri izdaji obveznic tudi tukaj upoštevati stroške izdaje (fl, *flotation costs*):

$$r_{ps} = \frac{D_{ps}}{P_o - fl}$$

5-5 Stroški navadnega lastniškega kapitala

Določitev stroškov navadnega lastniškega kapitala je bolj zapletena, saj so zahtevane donosnosti navadnih delničarjev odvisne od njihovih pričakovanj glede prihodnjega poslovanja podjetja. Glavno načelo pri določanju stroškov navadnega lastniškega kapitala je dejstvo, da morajo biti ti stroški oziroma zahtevane stopnje donosnosti enake kot jih navadni delničarji zahtevajo in dobijo z investicijami v podobno tvegane naložbe. Seveda moramo tudi tu upoštevati stroške izdaje navadnih delnic.

V nadaljevanju obravnavamo dve obliki navadnega lastniškega kapitala, zadržane dobičke in novoizdane navadne delnice.

5-6 Stroški zadržanih dobičkov

Ves dobiček po davkih, ki ostane podjetju, pripada lastnikom. Lahko ga vsega dobijo izplačanega v obliki dividend, lahko pa ga nekaj ostane v podjetju kot zadržani dobički. Zadržani dobički pomenijo za delničarje oportunitetni strošek v višini donosa, ki bi ga sicer lahko realizirali sami z naložbami s primerljivim tveganjem. Zato naj podjetje zadrži del dobička le, če bo z njim ustvarilo ustrezen donos.

Ključno za razumevanje je dejstvo, da zadržani dobički niso zastonj, ampak tako kot vsi drugi viri financiranja nekaj stanejo. Pri določanju stroškov zadržanih dobičkov je pomembna samo zahtevana donosnost lastniškega kapitala s strani navadnih delničarjev, saj ni stroškov izdaje (pridobitve) te vrste kapitala. Na zadržane dobičke lahko gledamo kot na izplačane dividende, ki jih lastniki takoj naložijo v nove navadne delnice podjetja.

Razvitih je bilo več metod za določanje stroškov zadržanih dobičkov, od katerih se največ uporabljajo metoda tržne zahtevane stopnje donosa, metoda premije za tveganje in pristop na podlagi CAPM modela. Omenjene tri metode vsebujejo subjektivne ocene določenih parametrov ali/in temeljijo na zelo močni predpostavkah, zato navadno dajo različne rezultate. Za končno oceno stroškov zadržanih dobičkov je ponavadi smiselno vzeti oceno CAPM pristopa (najbolj objektivna) in jo korigirati z drugima ocenama (ali več ocenami).

Metoda tržne zahtevane stopnje donosa

Metoda tržne zahtevane stopnje donosa (*dividend-yield plus growth rate model*) predpostavlja, da je za delnico podjetja poznana tržna cena (P_0) in da dividenda (D) ter cena delnice rasteta po konstantni stopnji (g). Zahtevana donosnost na delnico (r_s, indeks *s* pomeni *share*) je seštevek dividendne stopnje donosa (*dividend yield*) in pričakovane stopnje rasti cene delnice:

$$r_s = \frac{D_0(1+g)}{P_0} + g$$

Metoda premije za tveganje

Metoda premije za tveganje (*risk premium approach*) izhaja iz razlike med tveganjem upnikov in lastnikov podjetja. Lastniki podjetja nosijo večje tveganje, zato morajo biti tudi nagrajeni z ustrezno premijo za tveganje. Poleg poznavanja stroškov dolga pred davki (r_d) za podjetje je pri tej metodi ključna ocena razlike med donosnostjo dolga in navadnih delnic podjetja (*pr*, premija za tveganje):

$$r_s = r_d + rp$$

Metoda CAPM

Model določanja cen dolgoročnih naložb (CAPM, *Capital Asset Pricing Model*), ki smo ga spoznali v četrtem poglavju, omogoča izračun ocene zahtevane donosnosti delnice. Model izhaja iz netvegane naložbe oziroma donosnosti (r_f), ki ji doda tržno premijo za tveganje (r_m-r_f), pomnoženo s koeficientom β:

$$r_{s,i} = r_f + \beta_i \cdot (r_m - r_f)$$

5-7 Stroški novoizdanih navadnih delnic

Za razliko od zadržanih dobičkov ima podjetje pri izdaji novih navadnih delnic stroške izdaje, ki povečujejo strošek tega vira kapitala. Za izračun stroškov navadnih delnic se uporablja dopolnjena metoda tržne zahtevane donosnosti:

$$r_e = \frac{D_0(1+g)}{P_0(1-fl)} + g$$

Stroški izdaje (*fl*) so v zgornji enačbi izraženi kot delež cene delnice. Če pa so izraženi v absolutnem znesku, se enačba spremeni v

$$r_e = \frac{D_0(1+g)}{P_0 - fl} + g$$

5-8 Mejni stroški kapitala in točke preloma

Mejni stroški kapitala (MCC, *Marginal Cost of Capital*) so stroški zadnjega pridobljenega tolarja za financiranje nove dolgoročne naložbe. Podjetje zagotavlja nov kapital po WACC, toda ta strošek ni konstanten v nedogled. Slej ali prej naraste obseg dolga, navadnih in prednostnih delnic do meje, ker se njihovi stroški povečajo, posledično pa se poveča tudi WACC. Pravimo, da pride do točke preloma (BP, *Break Point*):

$$BP = \frac{\textit{obseg cenejšega kapitala}}{\textit{delež tega kapitala v vsem kapitalu}}$$

Vprašanja in naloge

V 5-1 Pri izračunu tehtanega povprečja stroškov kapitala (WACC) se uporablja stroške dolga pred davki.

a. Da b. Ne

V 5-2 Strošek zadržanih dobičkov je lahko manjši ali večji od stroška kapitala, zbranega z novo izdajo navadnih delnic, odvisno od stroškov izdaje in davčne stopnje.

a. Da b. Ne

V 5-3 Strošek kapitala odraža tehtani povprečni strošek različnih virov dolgoročnega financiranja, ki ga podjetje uporablja.

a. Da b. Ne

V 5-4 Podjetje mora zagotoviti donosnost novih projektov najmanj v višini mejnih stroškov kapitala.

a. Da b. Ne

V 5-5 Predpostavite **tipično** delniško podjetje. Katera od navedb glede stroška kapitala je točna (r_d pomeni strošek dolga, r_s strošek zadržanih dobičkov in r_e strošek navadnih delnic)?

a. $r_d > r_s > r_e >$ WACC
b. $r_e > r_s > r_d >$ WACC
c. WACC $> r_s > r_e > r_d$
d. $r_s > r_e >$ WACC $> r_d$
e. nobena od zgornjih navedb ni točna

V 5-6 V poslovnem časniku Finance ste zasledili oceno, da bodo dividende, čisti dobički in cena delnice podjetja Luka Poper d.d. naraščali po 8-odstotni letni stopnji. Prebrali ste tudi podatek, da je znašala zadnja izplačana dividenda 231,50 EUR in da se pričakuje naslednja dividenda v višini 250,00 EUR. V panogi, v katero sodi podjetje Luka Poper d.d., je povprečna razlika med stroški dolga in stroški navadnega lastniškega kapitala 4% točke. Luka Poper je nekoliko bolj tvegana kot je povprečno tveganje panoge, zato ji analitiki pripisujejo še dodatno odstotno točko k premiji za tveganje. Izračunajte strošek zadržanih dobičkov podjetja po spodaj navedenih metodah, če veste, da je trenutna cena delnice podjetja 3000 EUR.

a. Metoda tržne zahtevane stopnje donosa.
b. Metoda premije za tveganje, če je zahtevana donosnost obveznic podjetja 12%.

c. S pomočjo CAPM modela, če vemo, da je β podjetja 1,8, netvegana stopnja donosa 8% in povprečna stopnja donosa naložb na trgu 12%.
d. Kakšno ugotovitev glede stroška zadržanih dobičkov podjetja Luka Poper d.d. lahko sprejmemo na podlagi zgornjih izračunov?

V 5-7 Podjetje Optimizem d.d. ima naslednjo strukturo kapitala: dolg 50%, navadne delnice 30% in prednostne delnice 20%. Dobiček podjetja je 1 milijon EUR, od česar je 40% namenjenih za izplačilo dividend. Projekt kakšnega obsega lahko izpelje podjetje brez izdajanja novih navadnih delnic?

V 5-8 Delniška družba CTB d.d. želi nadaljevati s širitvijo svoje dejavnosti. Na podlagi analiz o segrevanju ozračja namerava v Ljubljani zgraditi prvo pokrito smučišče v Sloveniji. V podjetju pričakujejo velik obisk in zato napovedujejo pričakovano donosnost projekta v višini 13%. Podjetje bo tudi po izgradnji pokritega smučišča ohranilo trenutno razmerje med dolgom in lastniškim kapitalom D/E=0,7. Delnica podjetja trenutno kotira pri 4.200 EUR in pričakuje se izplačilo dividende v višini 250 EUR. Rast podjetje je ocenjena na 7%. Obstoječa serija obveznic (nominalna vrednost obveznice 100.000 EUR, letni kupon 8%, eno leto do dospetja) podjetja kotira po tečaju 96,5. Stopnja davka od dobička je 25%. Predpostavite, da želi CBT d.d. ohraniti obstoječo strukturo kapitala. Za izvedbo projekta ne bi bilo potrebno izdati novih delnic, za nov izdani dolg pa se zahteva enaka donosnost kot na že obstoječi dolg. Ali naj podjetje zgradi prvo slovensko pokrito smučišče?

V 5-9 Za rojstni dan se dobili delnico podjetja UniLj d.d., za katero v tečajnici Ljubljanske borze preberete trenutno ceno 550,00 EUR. Zasledite tudi podatek, da naj bi letošnji dobiček na delnico znašal 37,00 EUR in da ga bo izplačano 40% v obliki dividende. Kot novopečeni delničar zahtevate 12% donosnost na svojo delnico. Kakšna je pričakovana stopnja rasti podjetja UniLj d.d.?

V 5-10 Delnica se prodaja po 50 USD. Zadnja izplačana dividenda je bila 2 USD. Dividende rastejo po 6% letni stopnji rasti. Stroški nove izdaje delnic so 10%. Je strošek zadržanih dobičkov tega podjetja večji od stroška novih navadnih delnic?

V 5-11 Podjetje Privoz d.d. iz Novega mesta za prihodnje leto načrtuje 125 milijonov EUR čistega dobička, od katerega je 25% namenjeno izplačilu dividend. Optimalna struktura kapitala podjetja je 20% dolga in 80% lastniškega kapitala. Davčna stopnja je 25%, delnice podjetja kotirajo po 24.000 EUR, v prihodnosti pa se pričakuje 12% rast dividend (zadnje izplačana dividenda je znašala 1.300 EUR).

Podjetje razmišlja o novih dolgoročnih naložbah. Ocenili so, da so potencialno zanimive naložbe z delovnimi imeni A, B, C in D. Za te naložbe so znani naslednji podatki:

Naložba	Investicija v mio EUR	Pričakovana donosnost (%)
A	60	19,0
B	70	17,9
C	42	15,8
D	85	14

Če bi se podjetje odločilo, da nove projekte financira z izdajo novih delnic istega razreda, bi moralo plačati še 8% provizije od prodajne cene delnice. Poznani so tudi podatki o stroških zadolžitve pri banki:

Znesek posojila v mio EUR	Obrestna mera v %
do 30	16
od 30 do 47	18
nad 47	20

a. Odločitev, katere naložbe izpeljati, bo sprejeta s pomočjo izračuna mejnih stroškov kapitala (MCC), potrebnega za financiranje naložb. Koliko prelomih točk je na krivulji MCC in pri katerih zneskih kapitala pride do njih? Kakšno je tehtano povprečje stroškov kapitala (WACC) na vsakem od intervalov krivulje MCC?
b. Skicirajte krivuljo MCC in krivuljo investicijskih priložnosti (IOC).
c. Katere naložbe naj podjetje izpelje?

V 5-12 Optimalna struktura kapitala podjetja je 40% dolga in 60% lastniškega kapitala. Strošek dolga je 10%, dobiček podjetja je 40 milijonov EUR (za dividende je namenjeno 50% dobička). Davčna stopnja je 40%, cena delnice podjetja je 25.000 EUR, podjetje ima izdanih 10.000 delnic. Stopnja rasti podjetja je nič. Strošek nove izdaje delnic je 15%. Podjetje se pripravlja na naložbo v višini 40 milijonov EUR. Izračunajte tehtano povprečje stroškov kapitala (WACC) za omenjeni obseh naložb?

V 5-13 Podjetje financira svoje poslovanje s 40% dolga, 10% prednostnih delnic in 50% navadnega lastniškega kapitala. Dobiček podjetja je 100 milijonov USD (35% gre za izplačilo dividend). Obrestna mera dolga podjetja je 11%. Prednostne delnice se prodajajo po 20 USD in prinašajo dividendo 2 USD. Navadne delnice se prodajajo po 30 USD in imajo pričakovano dividendo 2 USD. Pričakuje se 8% rast podjetja. Stroški izdaje novih prednostnih delnic so 10%, navadnih pa 15%. WACC podjetja je pri načrtovani naložbi v višini 150 milijonov USD 12,43%. Kakšna je davčna stopnja?

Odgovori in rešitve

O 5-1 b.

O 5-2 b.

O 5-3 a.

O 5-4 a.

O 5-5 e.

O 5-6

a. 16,3%
b. 17%
c. 15,2%

O 5-7 2 milijona EUR

O 5-8 Da, saj je pričakovana donosnost projekta večja od stroška financiranja

O 5-9 9,3%

O 5-10 Ne (r_s=10,24% in r_e=10,71%)

O 5-11

b. BP1=117,2 milionov EUR (WACC1=16,86%), BP2=150 milionov EUR (17,28%), BP3=235 milionov EUR (17,6%).
c. Projekta A in B

O 5-12 8,05%

O 5-13 22,8%

6 OSNOVE INVESTICIJSKIH ODLOČITEV

Finančno načrtovanje dolgoročnih naložb je proces ugotavljanja in analize možnih dolgoročnih naložb. Odločitve o dolgoročnih naložbah so ene izmed najbolj zahtevnih odločitev, s katerimi se managerji srečujejo.

Finančno načrtovanje dolgoročnih naložb je pomembno, ker:
- se **učinki** dolgoročnih naložb **poznajo dalj časa v prihodnosti**, s čimer podjetje izgubi nekaj fleksibilnosti,
- so odločitve o dolgoročnih naložbah zasnovane na **pričakovanjih** o prihodnosti, ki je **negotova**,
- se podjetje z izborom dolgoročnih naložb strateško usmerja in pozicionira (trgi, proizvodi, …),
- ker je **doseganje cilja poslovanja podjetja** nenazadnje močno odvisno od učinkovitega investiranja v dolgoročne naložbe (investicijske odločitve na aktivi) ob optimalni strukturi financiranja podjetja (finančne odločitve na pasivi).

6-1 Razvrstitev in opredelitev projektov

- **NADOMESTITVENI PROJEKTI *(REPLACEMENT PROJECTS)***
  Vzdrževanje: nujno potrebni, če želi podjetje nadaljevati poslovanje v že doseženem obsegu.
  Znižanje stroškov: staro opremo, ki je še uporabna, vendar zastarela, nadomestimo z novo.

- **RAZŠIRITVENI PROJEKTI *(EXPANSION PROJECTS)***
  Obstoječi proizvodi in trgi: širitev na obstoječih trgih z/ali obstoječimi proizvodi; potrebne dobre napovedi rasti prodaje.
  Novi proizvodi in trgi: širitev na nove trge in/ali nove proizvode; strateške odločitve; zelo podrobna analiza.

- **RAZISKAVE IN RAZVOJ**

- **OSTALI PROJEKTI**
  Projekti, ki zadevajo uresničitev zahtev o varstvu pri delu, okoljevarstvenih zahtev, in vse ostalo (poslovne zgradbe, službeni avtomobili, itd.)

Finančno načrtovanje dolgoročnih naložb je podobno procesu vrednotenja finančnih naložb, saj je potrebno ravno tako oceniti vse prihodnje (pričakovane) denarne tokove in določiti ustrezno diskontno stopnjo (strošek kapitala) glede na tveganost pričakovanih denarnih tokov. Končno vse pričakovane denarne tokove prevedemo na sedanjo vrednost, kar nam služi kot osnova za primerjavo med naložbami in odločanje.

6-2 Investicijski kriteriji, njihove prednosti in slabosti

DOBA POVRAČILA (*PAYBACK PERIOD*)

Število let, v katerem se povrne začetni znesek naložbe brez upoštevanja časovne vrednosti denarja.

✓ Do določene mere lahko sklepamo o tveganosti in likvidnosti projekta
Enostavnost izračuna.

✗ Ne upošteva denarnih tokov za trenutkom povračila.
Ne upošteva časovne vrednosti denarja.

KRITERIJ: čim krajša doba povračila.

NETO SEDANJA VREDNOST (*NPV - NET PRESENT VALUE*)

Metoda ocenjevanja investicijskih projektov z uporabo tehnike diskontiranih denarnih tokov (*DCF- discounted cash flow*). Diskonta stopnja je enaka tehtanemu povprečju stroškov kapitala podjetja (*WACC - weighted average cost of capital*).

Postopek:
1. opredelimo vse pričakovane denarne tokove (*CF - cash flow*),
2. poiščemo sedanjo vrednost pričakovanih denarnih tokov - diskontiramo z WACC,
3. seštejemo sedanje vrednosti pričakovanih denarnih tokov in odštejemo začetni investicijski izdatek,
4. sprejmemo odločitev.

$$NPV = \sum_{t=0}^{n} \frac{CF_t}{(1+WACC)^t} = \sum_{t=1}^{n} \frac{CF_t}{(1+WACC)^t} - I_0$$

Odločitveni kriterij:
Če je NPV > 0, je investicijski projekt sprejemljiv.
Če je NPV = 0, je podjetje indiferentno do investicije (drugi kriteriji).
Če je NPV < 0, je investicijski projekt nesprejemljiv.

Izmed dveh medsebojno izključujočih projektov izberemo tistega z višjo NPV, če je NPV pozitivna, sicer oba zavrnemo.

✓ Upošteva vse pričakovane denarne tokove projekta.
Upošteva časovno vrednost denarja.

NOTRANJA STOPNJA DONOSA - IRR (*IRR - INTERNAL RATE OF RETURN*)

IRR je diskonta stopnja, pri kateri je sedanja vrednost pričakovanih denarnih pritokov projekta enaka sedanji vrednosti investicijskih izdatkov projekta oziroma je NPV projekta enaka 0.

$$\sum_{t=0}^{n} \frac{CF_t}{(1+IRR)^t} = 0$$

Odločitveni kriterij:
Če je IRR > WACC, je investicijski projekt sprejemljiv.
Če je IRR < WACC, je investicijski projekt nesprejemljiv.

NPV je sicer boljši kriterij pri odločanju o dolgoročnih naložbah kot IRR, vendar pa se IRR veliko uporablja v praksi, ker je relativna mera.

✓ Upošteva vse pričakovane denarne tokove projekta.
Relativna mera.
Informacija o varnostni meji projekta.

✗ Ni 100-odstotno zanesljiv kriterij za odločanje.

TEŽAVE Z NOTRANJO STOPNJO DONOSA PRI MEDSEBOJNO IZKLJUČUJOČIH PROJEKTIH

Kdaj?
Če se krivulji neto sedanje vrednosti projektov (na abscisi strošek kapitala, na ordinati NPV), ki ju primerjamo, sekata v zgornjem desnem kvadrantu koordinatnega sistema.

Zakaj?
- Različen obseg možnih naložb
- Različen časovni razpored denarnih tokov

Vzrok v IRR:
- Diskontiranje denarnih tokov z IRR, čeprav mora diskonta stopnja odražati zahtevano donosnost naložbe (strošek kapitala).
- Predpostavka o reinvestiranju denarnih prejemkov po IRR do konca življenjske dobe projekta.

Krivulja neto sedanje vrednosti nam prikazuje vpliv diskontne stopnje, po kateri diskontiramo denarne tokove projekta, na neto sedanjo vrednost projekta. Kjer krivulja seka abcisno os, je notranja stopnja donosa projekta, saj je neto sedanja vrednost projekta enaka nič.

PRIMER: Imamo dva medsebojno izključujoča projekta, projekt K (kratki) in projekt D (dolgi), z različnimi denarnimi tokovi. Pri projektu K nastopijo višji denarni tokovi v prvih letih projekta, pri projektu D pa bolj proti koncu življenjske dobe projekta.

Tabela 6-1: Denarni tokovi projektov K in D

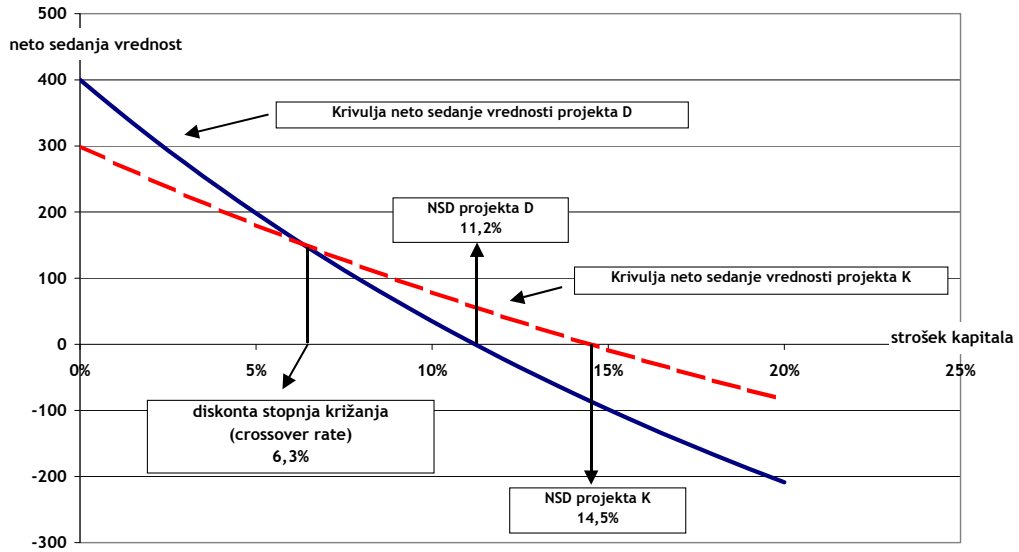
Leto	0	1	2	3	4
Projekt K	-1000	500	400	300	100
Projekt D	-1000	100	200	400	700

Iz slike 6-1 lahko sklepamo naslednje:

- Dokler je strošek kapitala večji od 6,3% (diskontna stopnja križanja), je NPV projekta K višja od NPV projekta D, kar pomeni, da po kriteriju NPV izberemo projekt K. IRR projekta K je višja od IRR projekta D, kar pomeni, da po kriteriju IRR izberemo projekt K. Dokler je strošek kapitala večji od 6,3%, je sklep o izbiri med projektoma K in D enak tako z uporabo NPV kot z uporabo IRR.
- Če je strošek kapitala manjši od 6,3%, je NPV projekta D višja od NPV projekta K, kar pomeni, da po kriteriju NPV izberemo projekt D. Vidimo pa, da po drugi strani po kriteriju IRR izberemo projekt K.

Kateri kriterij uporabimo? *Kriterij NPV je boljši, ker v vsakem primeru izberemo projekt z višjo NPV, kar pomeni, da bomo vedno izbrali tisti projekt, ki bo več prispeval k povečanju premoženja lastnikov.*

Slika 6-1: Krivulji neto sedanje vrednosti projektov K(ratki) in D(olgi)

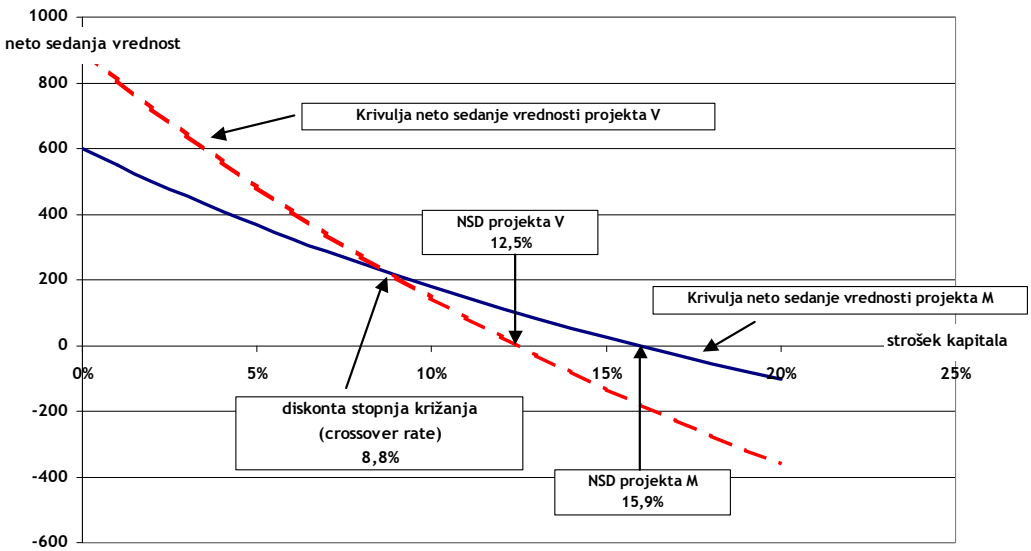


PRIMER: Imamo dva medsebojno izključujoča projekta, projekt V (veliki) in projekt M (mali), z različnimi denarnimi tokovi. Pri projektu V imamo višje investicijske izdatke in pritoke, pri projektu M pa nižje.

Tabela 6-2: Denarni tokovi projektov V in M

Leto	0	1	2	3	4
Projekt V	-2000	200	400	900	1400
Projekt M	-1000	100	200	500	800

Slika 6-2: Krivulji neto sedanje vrednosti projektov V(eliki) in M(ali)



Glede ustreznega investicijskega kriterija veljajo podobne ugotovitve kot pri prejšnjem primeru (projekta K in D). Zato bomo zaupali le NPV kot investicijskemu kriteriju, ki je med obravnavanimi najbolj ustrezen.

POPRAVLJENA NOTRANJA STOPNJA DONOSA (*MIRR - MODIFIED INTERNAL RATE OF RETURN*)

V praksi se je izkazalo, da imajo managerji raje kriterij, izražen z relativno mero (notranja stopnja donosa), kot pa kriterij, izražen z absolutno mero (NPV). Pri uporabi kriterija notranje stopnje donosa (IRR) lahko pride do vrste težav in ni najbolj zanesljiv kriterij. Popravljena notranja stopnja donosa (MIRR) te težave odpravlja.

Vprašanja in naloge

V 6-1 Ali je možno, da ima projekt dobo povračila 2 leti, neto sedanjo vrednost (NPV) pa negativno? Razložite.

V 6-2 Ali za neodvisne projekte velja, da je v primeru, ko je PI > 1, tudi NPV > 0 in IRR > WACC? Pokažite.

V 6-3 Razložite zakaj je NPV projekta z dolgo življenjsko dobo, kjer večji del denarnih tokov nastane proti koncu življenjske dobe projekta, bolj občutljiva na spremembe v stroških kapitala podjetja (WACC), kot pa NPV projekta s krajšo življenjsko dobo.

V 6-4 Začetni investicijski izdatek telekomunikacijskega projekta »A*synchronous Digital Subscriber Line*« (ADSL) znaša 165 milijonov EUR. Ob predpostavki, da bo projekt imel v naslednjih 10-ih letih, kolikor bo predvidoma trajal do prihoda nove tehnologije, 20.000 naročnikov, bo ustvarjal denarni pritok v višini 30 milijonov EUR letno. Strošek kapitala znaša 10%.

a. Kolikšna je doba povračila projekta? (PB)
b. Kolikšna je neto sedanja vrednost projekta? (NPV)
c. Kolikšna je notranja stopnja donosa projekta? (IRR)
d. Kolikšna je popravljena stopnja donosa projekta? (MIRR)
e. Kolikšen je indeks donosnosti projekta? (PI)

V 6-5 Ste finančni analitik podjetja Nalitika d.d. Direktor vas je zadolžil za analizo dveh potencialnih naložb (A in B), v kateri bi podjetje lahko investiralo. Naložba v vsakega izmed projektov stane 150 mio EUR, strošek kapitala pa je ocenjen na 15%. Pričakovani denarni tokovi obeh naložb, katerih življenjska doba je 4 leta, so naslednji:

Leto	Pričakovani denarni tok	
	Projekt A	Projekt B
0	-150	-150
1	90	55
2	45	55
3	45	55
4	30	55

podatki so v mio SIT

a. Izračunajte dobo povračila, neto sedanjo vrednost (NPV), notranjo stopnjo donosa (IRR) in popravljeno notranjo stopnjo donosa (MIRR) za vsakega izmed obeh projektov.

b. Kateri projekt (ali celo oba) bi sprejeli, če bi šlo za neodvisna projekta?
c. Kateri projekt bi sprejeli, če bi šlo za medsebojno izključujoča projekta?
d. Zakaj lahko pride do različnih odločitev o izbiri najboljšega projekta ob uporabi metode neto sedanje vrednosti (NPV) in uporabi metode notranje stopnje donosa? Skicirajte krivulji neto sedanje vrednosti (NPV) za oba projekta.
e. Ali bi se vaš odgovor na vprašanje c. spremenil, če bi znašal strošek kapitala 10%?

V 6-6 Ste finančni direktor podjetja Izpuh d.d. Pred nekaj meseci vam je inšpektorat za okolje izrekel negativno mnenje glede vaše proizvodnje, saj le-ta preveč onesnažuje okolje, in vam dal pol leta časa, da izboljšate kakovost zraka, ki prihaja iz dimnikov vaše tovarne. Zato je ekipa strokovnjakov iz vašega podjetja pripravila predlog za vgradnjo čistilnega filtra, ki bi odpravil pomanjkljivosti glede onesnaževanja v vaši proizvodnji. Vgradnja čistilnega filtra nikakor ne vpliva na zmogljivosti vaše proizvodnje oziroma pričakovane prihodke prodaje[14]. Na podlagi razpisa ste pridobili ponudbi dveh proizvajalcev filtrov, katerih življenjska doba je 5 let. Strošek kapitala podjetja je 12%.

Leto	Pričakovani neto denarni tok	
	Filter A	Filter B
0	-60000	-80000
1	-15000	-6000
2	-15000	-6000
3	-15000	-6000
4	-15000	-6000
5	-15000	-6000

podatki so v 1000 SIT

a. Izračunajte notranjo stopnjo donosa (IRR) za vsako izmed obeh možnosti.
b. Koliko znaša neto sedanja vrednost stroškov obeh možnosti? Katero ponudbo za namestitev filtrirne naprave boste izbrali?

[14] dejstvo, da bi morali tovarno v nasprotnem primeru zapreti (celoten izpad prihodkov), bomo zaradi poenostavitve primera odmislili.

Odgovori in rešitve nalog

O 6-1 Da, odvisno od diskonte stopnje in časovnega razporeda denarnih pritokov.

O 6-2 Da.

O 6-4

a. 5,5 let.
b. 19,33 mio EUR.
c. 12,7%.
d. 11,2%.
e. 1,117.

O 6-5

a.

NPV	9,03	7,02
IRR	18,6%	17,3%
MIRR	16,7%	16,3%
PB	2,33	2,73

b. oba projekta.
c. projekt A.
d. zaradi pomanjkljivosti metode IRR.
e. da, izbrali bi projekt B.

O 6-6

a. izračun ni možen.
b. NPV projekta Filter A znaša -114.071,6, NPV projekta Filter B pa -101.628,7. Izbrali bi projekt Filter B, ker ima po absolutni vrednosti nižjo NPV.