# MODULE I

# 1

# INTRODUCTION TO GAME THEORY

**Unit Structure**





## 1.0 OBJECTIVES

- To understand the various concepts of games.
- To know the meaning of Prisoner's Dilemma.
- To study the concept of duopoly price war.

## 1.1 INTRODUCTION

In a climate of uncertainty, economic decision making involves strategy. Every firm needs to find out as to how other firms will react to price and output decisions. Will there be a price war and if so, would it lead to losses. Will bargaining with the workers union would end in a stalemate and strike. The making of the Union budget involves a lot of bargaining between the various stake holders in the society. The trade unions, associations of commerce and industry, consumer groups, political parties and other interest groups get involved in influencing the budget. The study of economic games that these stake holders play is known as Game Theory. Economic decision making thus involves uncertainty and strategy.

Game theory is an important branch of economic theory and analysis that provide many insights into the behavior of economic agents in situations where there is an *actual or potential conflict of interest.* It is an approach to analyzing rational decision making behavior in interactive or conflict situations. Game theory analyses the way that two or more players or parties choose actions or strategies that jointly affect each participant. The element of game arises because the outcome depends not only on the choices made by one player but also on what other players choose to do at the same time. **This theory was developed by John von Neumann (1903-57) and Oskar Morgenstern in their work "The Theory of Games and Economic Behavior".** Game theory has been used by economists to study the interaction of duopoly, monopolistic and oligopoly firms, union management disputes, trade policies etc.

## 1.2 BASIC CONCEPTS

According to **Walter Nicholson**, a game is a situation in which individuals must make decisions and in which final outcome will depend on what each person decides to do. In a game, agents aim to maximize their own pay-off by choosing specific actions but the actual outcome depends on what all other players do. The game consists of a specified interactive playing field, a specification of all possible courses of action and a schedule of the pay-offs to each of the players under all possible outcomes. Players plan their own courses of action in order to maximize their expected payoff, under the knowledge that the other players are trying to do the same. **Any game has three basic elements. They are: the players, the list of possible actions or strategies available to each player and the payoffs the players receive for each possible combination of strategies.** The player in the game theory is the decision maker. Firms are considered to be players in oligopoly markets. The number of players is generally fixed throughout the game and some games have a fixed number of players. Strategy refers to a course of action available to a player in a game. Generally players do not have too many options as far as strategy is concerned. Payoffs refer to final outcomes to the players at the end of the game.

A player's strategy is a complete specification of the actions to be taken in response to outcomes that are found as the game proceeds. A player's pay-off from choosing a strategy depends on what the other players do but players cannot make binding agreements with each other. Given the strategies of all the players, there will be a set of possible outcomes to the game. These determine the potential pay-offs for each of the players. **A specific outcome is called equilibrium if no player can take actions to improve his own pay-off while all other players continue to follow their optimal strategies.** In order to select the best strategy, a player must know what other players will do but they in turn must also know every player will do. In strategic game, players choose their moves simultaneously. Whenever the choices are discrete and finite, the game

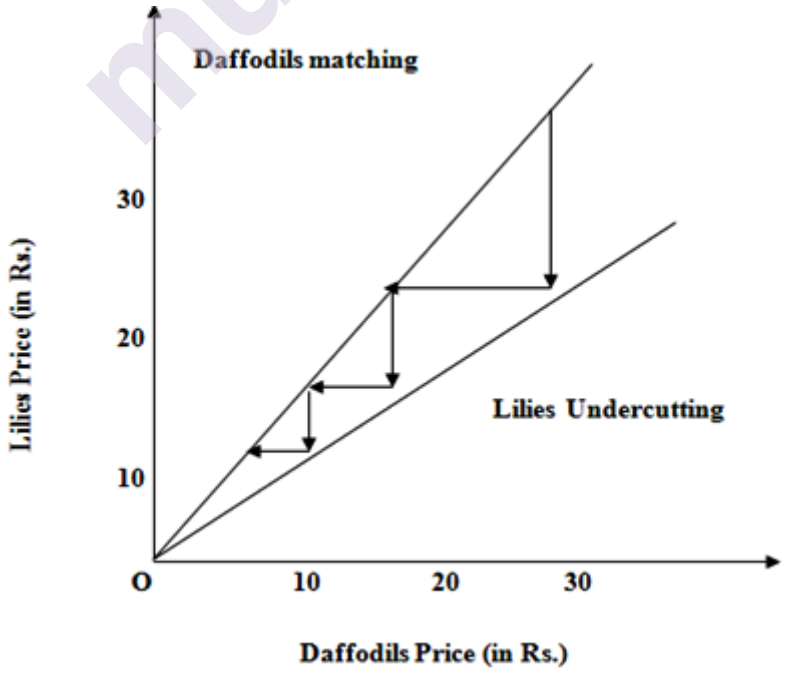
can be represented in the structure of a table which sets out the outcomes for each player depending on what the other players do. In an extensive game, players make moves in some order and hence the analysis of the game needs a specification of the pay-offs and information at each point in time. Real business interactions are similar to an extensive game, as firms interact dynamically over a period of time. However, whenever the precise timing of moves is not essential to the outcome, a game can be represented as a normal game. A game that is played only once is a 'one-shot-game'. Repeated games open possibilities of learning and of acting in order to punish or reward the other players. A super-game is a game that is repeated many times.

The basic concepts of Game theory are being explained by studying a duopoly price war.

## 1.3 DUOPOLY PRICE WAR

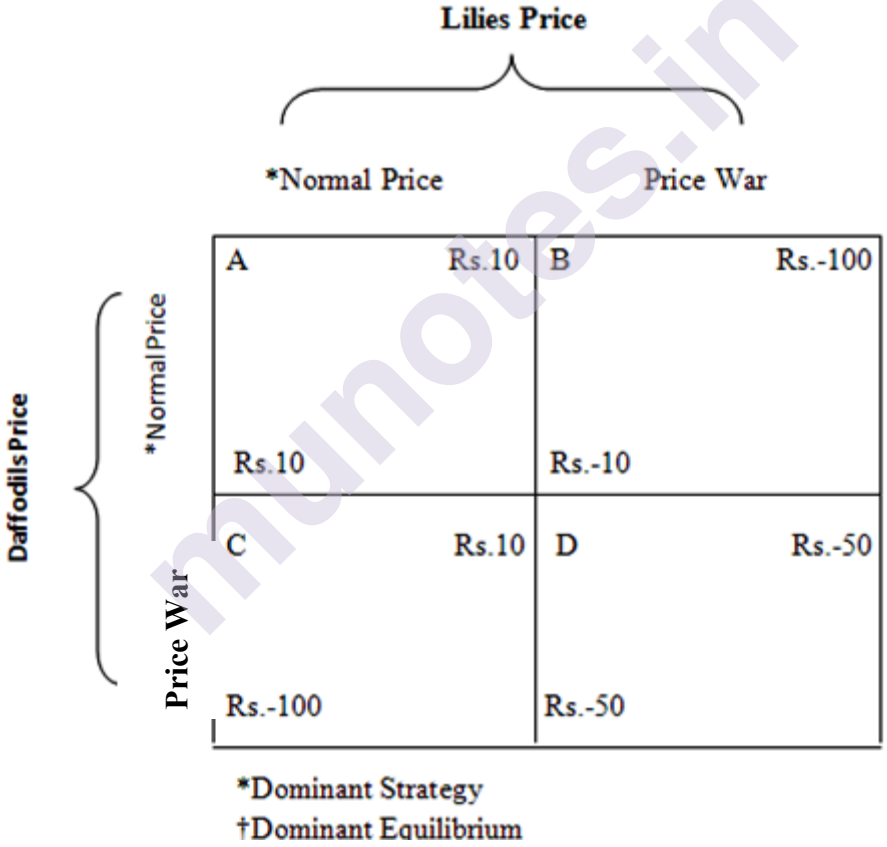
Let us assume that you are the head of Daffodils, a departmental store whose motto is "We will not be undersold". Your rival Lilies, runs an advertisement, "We sell for ten per cent less". Figure 1.1 shows the dynamics of price cutting. The vertical arrows shows Lilies price cuts, the horizontal arrows shows Daffodils responding strategy of matching each price cut. Notice that the pattern of reaction and counter-reaction will end up in a zero price because the only price compatible with both strategies is a zero price. Lilies ultimately realize that when it cuts its price, Daffodils will match the price cuts. Now you will begin to ask what Lilies will do if you charge price A, B, C etc. Once you begin to consider how others will react to your actions, you have entered the arena of Game Theory.

**Fig.1.1 - Price War**



Duopoly is a situation where the market is supplied by two firms that are deciding whether to engage in price war and destroy themselves. Let us assume for the sake of simplicity that both the firms have the same cost and demand structure. Further, each firm can choose whether to charge its normal price or lower its price below the marginal costs and drive away the rival. In this duopoly game, the firm's profits will depend on its strategy and that of its rival's. The interaction between the two firms or people is represented by a two-way pay-off table. A pay-off table is a means of showing the strategies and the pay-offs of a game between two players. Figure 1.2 shows the pay-offs in the duopoly price game for our two stores. In the pay-off table, a firm can choose between the strategies listed in its rows or columns. For example, Lilies can choose between its two columns and Daffodils can choose between its two rows. Here, each firm decides whether to charge its normal price or to begin a price war by choosing a low price.

**Fig. 1.2 – A Pay-off Table for Price War.**




By combining the two decisions of each Duopoly firm gives four possible outcomes which are shown in the four cells of the table. The number in the lower left shows the pay-off to Daffodils and the numbers in the upper right shows the pay-off to Lilies.

## 1.4 ALTERNATIVE STRATEGIES

In Game theory, you are required to think through the goals and actions of your opponent and to make your decisions based on your opponent's goals and actions. While you think through your opponents, you must remember that your opponent will also be trying to outwit you. The following is the guiding philosophy in Game theory:

> **"Choose your strategy by asking what makes most sense for you assuming your opponent is analyzing your strategy and acting in his best interest."**

Let us apply this philosophy to the Duopoly example. Note that both the firms have the highest joint profits in outcome A. Each firm earns Rs.10 when both follow a normal price strategy. At the other end is price war where each cuts prices and runs a big loss. Between the two extreme ends, there are two interesting strategies where only one firm engages in price war. In outcome C, Lilies follow a normal price strategy while Daffodils engages in a price war. Daffodils take away most of the market but makes heavy losses because it is selling below cost. Lilies' is better-off selling at normal prices rather than responding and as a result, his loss is only Rs.10 against a loss of Rs.100 made by Daffodils.

## 1.5 DOMINANT STRATEGY

To begin with the game, one must know whether each player has a dominant strategy. This situation arises when one player has a best strategy no matter what strategy the other player follows. In the price war game example, consider the options open to Daffodils. If Lilies conducts business as usual with a normal price, Daffodil will get Rs.10 profit if it plays the normal price and will lose Rs.100 if it declares price war. On the other hand, if Lilies starts a price war, Daffodils will lose Rs.10 if it follows the normal price but will lose more if it also engages in price war i.e. Rs.50. The same logic holds true for Lilies. Therefore, no matter what strategy the other firm follows, each firm's best strategy is to have the normal price. Charging the normal price is a dominant strategy for both firms in the price war game.

A strategy is a dominant strategy for a player if its outcome or pay-off is most favorable given the available alternative strategies irrespective of what his competitor does. When both players have a dominant strategy, we say that the outcome is a dominant equilibrium. In Figure 1.2 above, outcome A is a dominant equilibrium because it arises from a situation where both firms are playing their dominant strategies.

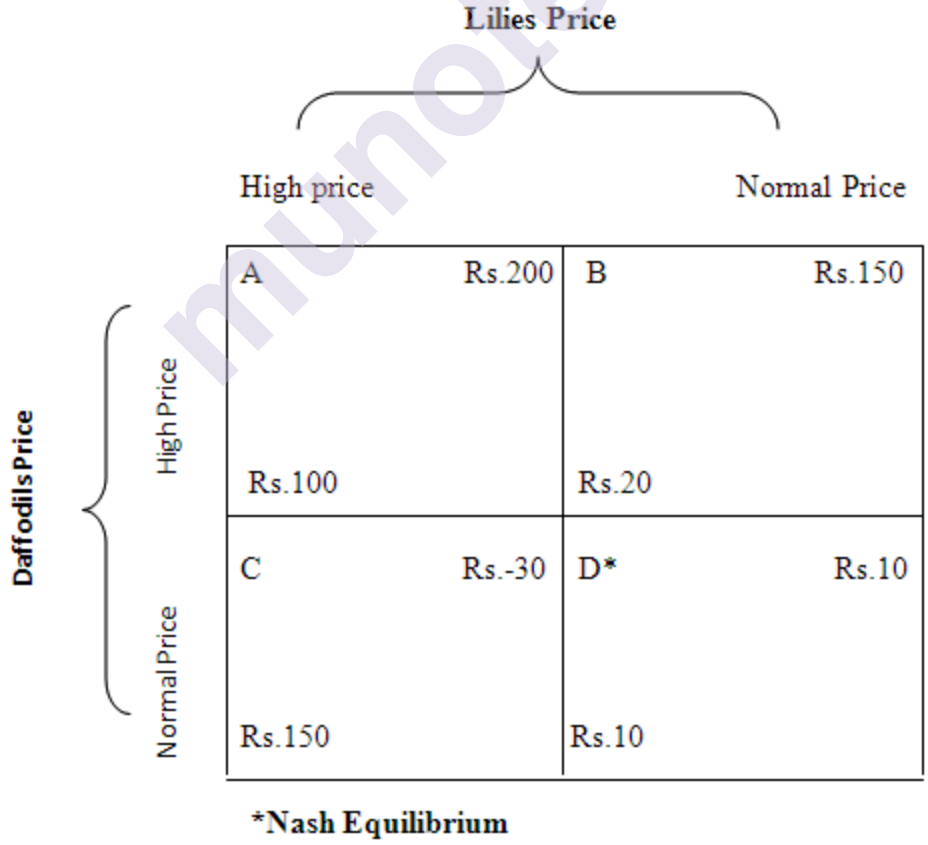
## 1.6 NASH EQUILIBRIUM

Most interesting situations do not have a dominant equilibrium. We can use our duopoly example to find this out. In a game of rivalry, each firm

considers whether to have its normal price or a monopoly price and earn monopoly profits. The game of rivalry is shown in Figure 1.3.

The firms can decide on the normal price equilibrium as found in the price war game or they can raise their price to earn monopoly profits. Notice that both the firms have the highest joint profits in Cell 'A' where they can earn a total of Rs.300 when each follows a high price strategy. Situation 'A' can emerge if the firms collude and set the monopoly price. At the other extreme is the competitive strategy of normal price where each rival has profits of only Rs.10. In between the two extremes there are two interesting strategies where one firm chooses a normal price and the other one a high price strategy. In Cell 'C', Lilies follows a high price strategy but Daffodils' undercuts. Daffodils' take away most of the market and has the highest profit from any of the four situations and Lilies loses money. In Cell 'B', Daffodils gambles on high price but Lilies normal price means a loss for Daffodils. In this example, Daffodils has a dominant strategy. It will profit more by choosing a normal price no matter what Lilies does. On the other hand, Lilies does not have a dominant strategy because Lilies would want to play normal if Daffodils plays normal and would want to play high if Daffodils play high. Lilies' has an interesting dilemma. Should it play high and hope that Daffodils will follow or play safe by playing normal.

**Fig 1.3 The Game of Rivalry (Should a Duopoly Try the monopoly price)**




***Nash Equilibrium**

By thinking through the pay-offs, it becomes clear that Lilies should play the normal price. The reason is that Lilies should start by putting itself in Daffodils' shoes. Notice that Daffodils' will play normal price no matter what Lilies does because that is Daffodils dominant strategy. Therefore Lilies should find its best action by assuming that Daffodils will follow his best strategy which immediately leads to Lilies playing normal. This illustrates the basic rule of Game theory: **"You should set your strategy on the assumption that your opponent will act in his best interest."**

The solution is called the **Nash equilibrium** after mathematician John Nash who developed the concept in the 1950s and won the Nobel Prize in Economics in 1994 for his contributions to the Game theory. **A Nash equilibrium is one in which no player can improve his or her pay-off given the other player's strategy. That is, given player 'A's strategy, player 'B' can do no better and vice-versa. Each strategy is a best response against the other player's strategy. The Nash equilibrium is called the non-cooperative equilibrium because each party chooses that strategy which is best for itself without collusion or co-operation and without regard for the welfare of society or any other party.** According to Nash theorem, every game with a definite number of players and a definite number of strategies would at least have one 'Nash equilibrium'. However, in order to hold the Nash theorem to be true, the strategies available must have some random element to them. A strategy with some random element is known as a mixed strategy. There may be multiple Nash equilibrium and it may not be clear as to which one will arise. Further, it is generally true that the Nash equilibrium is not the global optimum i.e. if players could co-operate they could all become better off. A game theory framework can help us understand the strategic choices available but it does not always help predict which of many possible outcomes may occur.

**Nash Equilibrium In Pure Strategies:**
When a player adopts a single strategy and holds on to it, it is known as a pure strategy. However, if the player uses two or more strategies in order to keep his opponents guessing, it is known as a mixed strategy. The situation of a pure strategy is deterministic because there is no change in the strategy whereas in the case of a mixed strategy, it is probabilistic because each strategy from the bundle has a probability of being picked up. On the basis of total gain or total loss, games are classified into zero-sum and non-zero-sum games. When there are two competitors in a game, it is called a two-player game and when the number of player are more than two, it is called a n-player game.

## 1.7 A ZERO-SUM GAME

In a zero-sum game, one man's gain is equal to another man's loss i.e. the sum of a positive number (gain) and that of a negative number (loss) is equal to zero. The Pay-off Matrix in a zero-sum game of Local Body

Elections is presented in Figure 1.4. Figures in the cells shows pay-offs to the two candidates Anil and Sunil. Positive signs show gains whereas negative signs show losses. These two candidates have two strategies i.e. Social Workers campaigning or Businessmen campaigning for their elections. If both the candidates use Social Workers for their campaigning, the outcome is zero i.e. nobody gains or loose as shown in Cell 'A'. The top right cell or Cell 'B' in the matrix shows that Anil's pay-off or gain of votes is 3000 with his strategy of using social workers for the campaign as against Sunil's strategy of using Businessmen. The bottom left and right cells i.e. Cells 'C' and 'D' indicate that Anil's pay-off for using the strategy of Businessmen is -2000 and -1000 i.e. loss of 3000 votes to Sunil for his strategy of using social workers and businessmen. In this game, gains made by Sunil are at the expense of Anil i.e. (+3000) + (-3000) = 0. In this game, Cell 'A' shows the dominant strategy equilibrium.

**Figure No. 1.4 – The Payoff Matrix for a Local Body Elections (Zero-sum Game)**

| | | **Sunil** | |
|---|---|---|---|
| | | **Social Workers Campaign** | **Businessmen Campaign** |
| **Anil** | **Social Workers Campaign** | A<br>0 | B<br>+3000 |
| | **Businessmen Campaign** | C<br>-2000 | D<br>-1000 |



## 1.8 A NON-ZERO-SUM GAME

In a non-zero-sum game, the sum of one player's gain and the loss of the competitor is non-zero. In the Oligopolistic game played in Figure 1.5, the outcome is a non-zero-sum outcome. There are two firms, namely: Daffodils and Lilies. Both the firms have two strategies i.e. normal price and monopoly price. If either firm follows a monopoly price strategy, both gains better pay-offs i.e. from Rs.100 to Rs.600 each, shown in cell 'D' or the bottom right cell. Cell 'D', however, indicates collusive or cooperative

equilibrium. If one of the firms maintains normal price and the other a high price, the one who follows a normal price strategy gains Rs.900 i.e. from Rs.100 to Rs.1000 whereas the one who follow a high price strategy suffers a loss of Rs.60 i.e. from Rs.100 to Rs.40 as shown in Cell 'B'. The sum of the changes in the pay-offs of the two firms is non-zero because the total profit to be earned is not fixed like the total number of votes in the game of Local Body elections. Cell 'A' shows that both the firms have a dominant strategy to charge a normal price and earn Rs.100 each. Cell 'A' also shows Nash Equilibrium given its definition because the strategies are reciprocal and the pay-offs are equal.

As against a game of rivalry, in a cooperative or collusive game, the players are assumed to be rational to understand that their mutual interest is in cooperation and not in competition. In a game of cooperation, at least one of the players will benefit without causing a loss to the other. In a non-cooperative game or a game involving rivalry, the players do not cooperate with each other for want of communication. The prisoners' dilemma explained in Table 1.3 is an example of non-cooperative game.

**Figure No. 1.5 – The Payoff Matrix in a Game of Rivalry (Non-zero-sum Game)**

| | | **Lilies** | |
|---|---|---|---|
| | | **Normal Price** | **High Price** |
| **Daffodils** | **Normal Price** | A* Rs.100<br>Rs.100 | B Rs.40<br>Rs.1000 |
| | **High Price** | C Rs.1000<br>Rs.40 | D Rs.600<br>Rs.600 |

## 1.9 PRISONER'S DILEMMA

In the prisoner's dilemma, when each player chooses his dominant strategy, the result is unfavorable to both the players. There are two

prisoners, Anil and Sunil who are locked up in separate cells for committing a crime. However, the prosecutor has limited hard evidence to convict them for a minor offence for which the punishment is one year imprisonment. Each prisoner is told that if one admits while the other remains silent, the confessor will be let off without being imprisoned and other one will be jailed for 20 years. If both the prisoners confess, they will be jailed for only five years. The two prisoners are not allowed to communicate with each other. The payoffs to the prisoners are shown in Figure 1.6.

**Figure No. 1.6 – The Payoff Matrix for a Prisoner's Dilemma**

| | | **Anil** | |
|---|---|---|---|
| | | **Confess** | **Remain Silent** |
| **Sunil** | **Confess** | A<br><br>5 Years for each | B<br><br>Zero years for Sunil<br>20 years for Anil |
| | **Remain Silent** | C<br><br>20 years for Sunil<br>Zero years for Anil | D<br><br>1 year for each |



In this game, the dominant strategy for both the players is to confess irrespective of the strategy pursued by the other as shown in Cell 'A'. Irrespective of Anil's strategy, Sunil will get a lighter sentence by confessing. If Anil admits to the crime, Sunil will get five years (Cell 'A') instead of 20 (Cell 'C'). If Anil remains silent, Sunil will be let off (Cell 'B') instead of spending a year in jail (Cell 'D'). As the payoffs are perfectly symmetric, Anil will also be happy to confess irrespective of what Sunil does. The difficulty is that when each follows his dominant strategy and confesses, both will do worse than if each had shown restraint. When both confesses, each get five years (Cell 'A') instead of the one year they would have gotten by remaining silent (Cell 'D'). The choices before the prisoner's exemplify a dilemma in which the prisoners have to make a choice between two evils i.e. to confess or to remain silent.

Oligopoly firms face similar dilemma when they introduce monopoly price as against the price set by the rival. The oligopoly firms need to decide as to whether they should compete or cooperate. There is however difference in the situation explained by prisoners' dilemma and the oligopoly situation. The prisoners here have only one chance to choose a strategy whereas the oligopoly firms have more than one chance to choose

their strategies. They have the opportunity, learn, unlearn and relearn and hence the probability of collusion or cooperation is high amongst oligopoly firms. Oligopoly firms may also decide to compete and get involved in a price war in order to increase their market shares. However, the existence of a kinked demand curve only proves the existence of price rigidity or price stability in the oligopoly market. Further, the existence of Carters and Price Leader firms and price signaling mechanism only proves that there is a desire for stability amongst the oligopoly firms.

## 1.10 NORMAL FORM GAME.

A normal form game or a n-player game is any list G = (S1,...,Sn; u1,...,un), where, for each i ∈ N = {1,...,n}, Si is the set of all strategies that are available to player i, and ui : S1 × ... × Sn → R is player i's von Neumann-Morgenstern utility function. A player's utility depends not only on his own strategy but also on the strategies played by other players. Moreover, ui is a von Neumann-Morgenstern utility function so that player 'i' tries to maximize the expected value of ui (where the expected values are computed with respect to his own beliefs). Here, player i is rational if he tries to maximize the expected value of ui (given his beliefs). It is also assumed that it is common knowledge that the players are N = {1,...,n}, that the set of strategies available to each player i is Si, and that each i tries to maximize expected value of ui given his beliefs. When there are only 2 players, we can represent the (normal form) game by a bi-matrix as shown below.

| 1\2 | left | right |
|---|---|---|
| up | 0,2 | 1,1 |
| down | 4,1 | 3,2 |

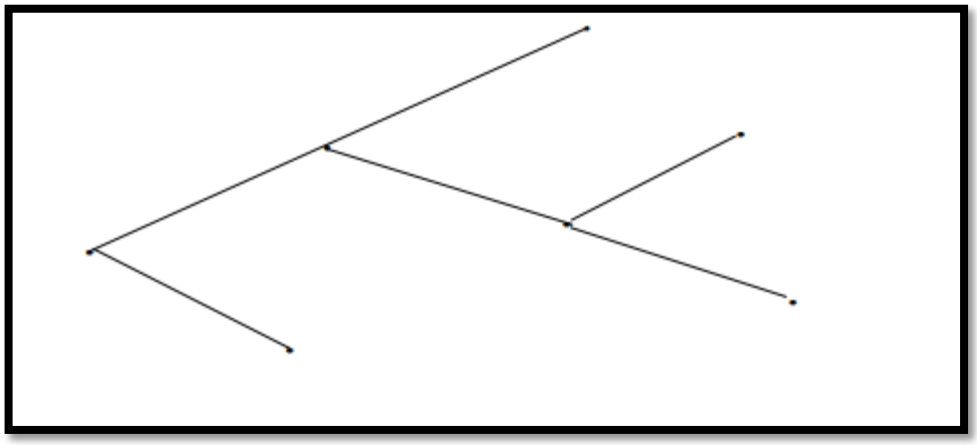
Here, Player 1 has strategies up and down, and player 2 has the strategies left and right. In each box the first number is 1's payoff and the second one is 2's (e.g., u1 (up, left) = 0, u2 (up, left) = 2).

## 1.11 EXTENSIVE FORM GAMES

The extensive form contains all the information about a game, by defining who moves when, what each player knows when he moves, what moves are available to him, and where each move leads to, etc.

A tree is a set of nodes and directed edges connecting these nodes such that 1) for each node, there is at most one incoming edge; 2) for any two nodes, there is a unique path that connects these two nodes. Imagine the branches of a tree arising from the trunk. For example, Figure 1.7 is a tree.

**Figure No. 1.7**



However, figure 1.8 shown below is not a tree because there are two alternative paths through which point A can be reached (through B and through C). So also, the figure is not a tree either since A and B are not connected to C and D.

**Figure No. 1.8**

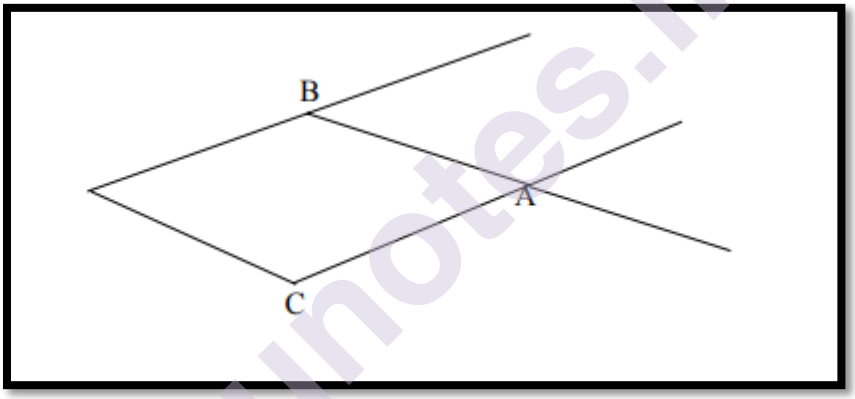
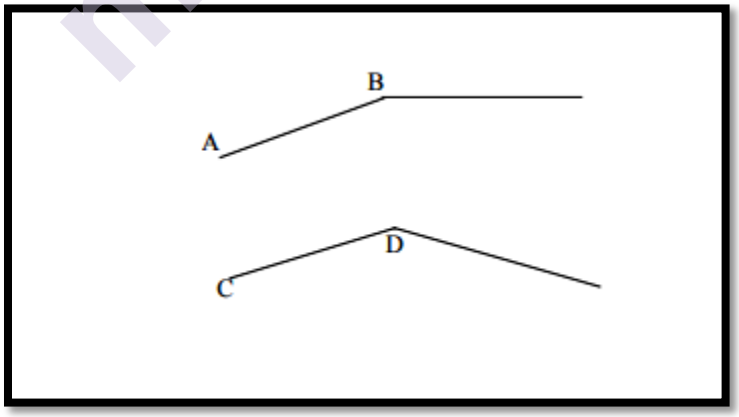



**Figure No. 1.9**




An extensive form Game consists of a set of players, a tree, an allocation of each node of the tree (except the end nodes) to a player, an informational partition, and payoffs for each player at each end node. The set of players will include the agents taking part in the game. However, in many games there is room for chance, e.g. the throw of dice in

backgammon or the card draws in poker. One need to consider the "chance" whenever there is uncertainty about some relevant fact. To represent these possibilities a fictional player such as Nature is introduced. There is no payoff for Nature at end nodes, and every time a node is allocated to Nature, a probability distribution over the branches that follow needs to be specified, e.g., Tail with probability of 1/2 and Head with probability of 1/2. An information set is a collection of points (nodes) {n1,...,nk} such that 1) the same player i is to move at each of these nodes; 2) the same moves are available at each of these nodes. Here the player i, who is to move at the information set, is assumed to be unable to distinguish between the points in the information set, but able to distinguish between the points outside the information set from those in it. For instance, consider the game in Figure 1.10. Here, Player 2 knows that Player 1 has taken action T or B and not action X; but Player 2 cannot know for sure whether 1 has taken T or B. The same game is depicted in Figure 1.11 slightly differently. An information partition is an allocation of each node of the tree (except the starting and end-nodes) to an information set.

**Figure No. 1.10**

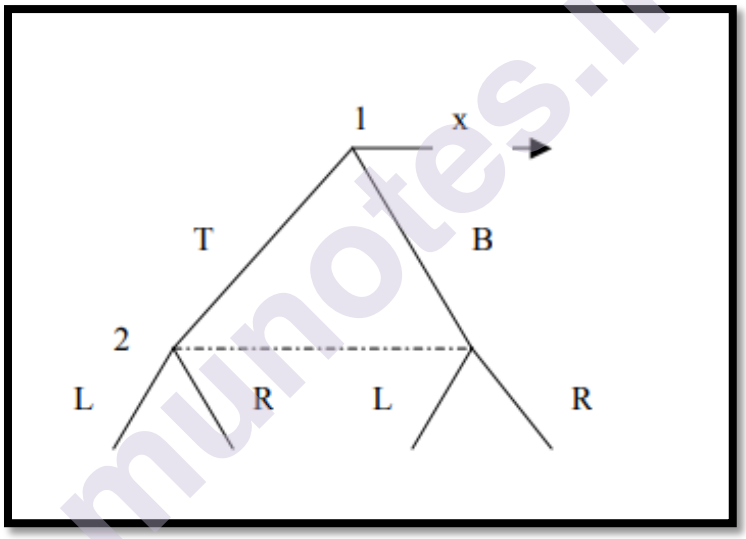
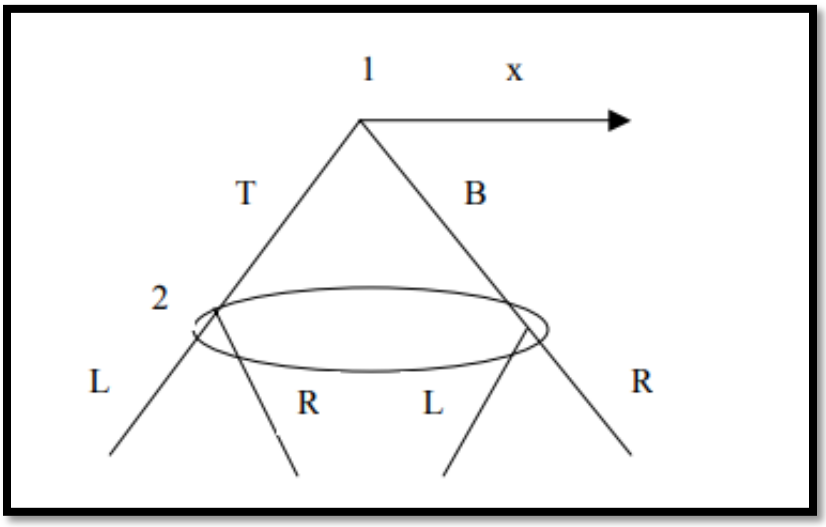



**Figure No. 1.11**



To conclude, at any node, it is well known as to which player is to move, which moves are available to the player, and which information set contains the node, summarizing the player's information at the node. If two nodes are in the same information set, the available moves in these nodes must be the same, for otherwise the player could distinguish the nodes by the available choices. And all these are assumed to be common knowledge. For instance, in the game in Figure 1.10, player 1 knows that, if he takes X, player 2 will know this, but if he takes T or B, player 2 will not know which of these two actions has been taken. (he will know that either T or B will have been taken).

## 1.12 SUB-GAME PERFECTION.

A smaller game that is part of an extensive form game is called a sub-game. When backward induction is restricted to a sub-game, the equilibrium computed for the main game, remains equilibrium for the sub-game also. Sub-game perfection generalizes this idea to general dynamic games. Nash equilibrium is said to be sub-game perfect if it is so in every sub-game of the game. A sub-game must be a well defined game when it is considered individually. The sub-game must have an initial node and all the moves and information sets from that node must remain in the sub-game.

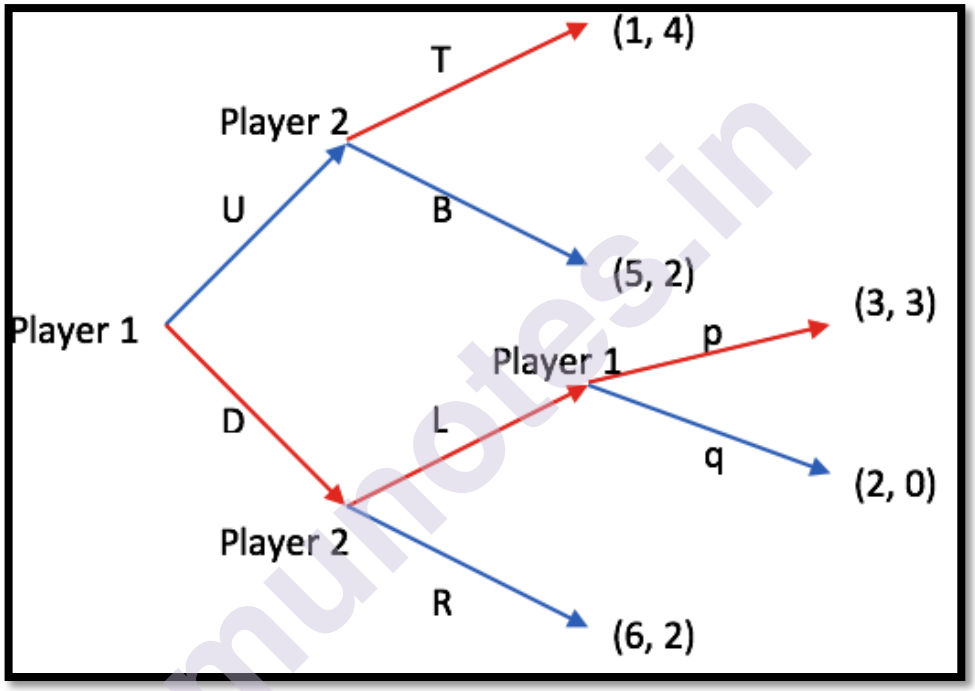
In game theory, a sub-game perfect equilibrium is a refinement of a Nash equilibrium used in dynamic games. A strategy profile is a sub-game perfect equilibrium if it represents a Nash equilibrium of every sub-game of the original game. Informally, this means that at any point in the game, the players' behavior from that point onward should represent a Nash equilibrium of the continuation game (i.e. of the sub-game), no matter what happened before. Every finite extensive game with perfect recall has a sub-game perfect equilibrium. Perfect recall is a term introduced by Harold W. Kuhn in 1953 and "equivalent to the assertion that each player is allowed by the rules of the game to remember everything he knew at previous moves and all of his choices at those moves".

A common method for determining sub-game perfect equilibrium in the case of a finite game is backward induction. Backward induction is the process of reasoning backwards in time, from the end of a problem or situation, to determine a sequence of optimal actions. It proceeds by examining the last point at which a decision is to be made and then identifying what action would be most optimal at that moment. Using this information, one can then determine what to do at the second-to-last time of decision. This process continues backwards until one has determined the best action for every possible situation (i.e. for every possible information set) at every point in time. Backward induction was first used in 1875 by Arthur Cayley, a British Mathematician. Here one first considers the last actions of the game and determines which actions the final mover should take in each possible circumstance to maximize his/her utility. One then supposes that the last actor will do these actions, and

considers the second to last actions, again choosing those that maximize that actor's utility. This process continues until one reaches the first move of the game. The strategies which remain are the set of all sub-game perfect equilibrium for finite-horizon extensive games of perfect information. However, backward induction cannot be applied to games of imperfect or incomplete information because this entails cutting through non-singleton information sets.

For example, determining the sub-game perfect equilibrium by using backward induction is shown below in Figure 1.12. Strategies for Player 1 are given by {Up, Uq, Dp, Dq}, whereas Player 2 has the strategies among {TL, TR, BL, BR}. There are four sub-games in this example, with 3 proper sub-games.

**Figure No. 1.12**




**A Sub-game Perfect Equilibrium.**

Using the backward induction, the players will take the following actions for each sub-game:

1. Sub-game for actions p and q: Player 1 will take action p with payoff (3, 3) to maximize Player 1's payoff, so the payoff for action L becomes (3,3).
2. Sub-game for actions L and R: Player 2 will take action L for 3 > 2, so the payoff for action D becomes (3, 3).
3. Sub-game for actions T and B: Player 2 will take action T to maximize Player 2's payoff, so the payoff for action U becomes (1, 4).
4. Sub-game for actions U and D: Player 1 will take action D to maximize Player 1's payoff.

Thus, the sub-game perfect equilibrium is {Dp, TL} with the payoff (3, 3).

## 1.13 QUESTIONS

Q1. Write a note on Prisoners' Dilemma.

Q2. Explain how Nash equilibrium is achieved in pure and mixed strategies.

Q3. Write a note on normal and extensive form games.

Q4. Write a note on Sub-game Perfection.

*****

# 2

# RISK AND UNCERTAINTY

**Unit Structure**



## 2.0 OBJECTIVES

To learn and understand the concept of risk in brief and the concept of uncertainty in details. In this we will study the behaviour of a rational consumer under uncertainty and how he makes a choice under uncertainty. You will study the behaviour of risk-avers, risk-neutral and risk-loving people. You will learn to study the use of tables, equations and graphs to study this behaviour

## 2.1 INTRODUCTION

Many of the choices that people make involve considerable uncertainty. Sometimes we need to choose between risky ventures. For example, what should we do with our savings? Should we invest in something safe, such as a bank savings account, or something riskier but more lucrative, such as the stock markets? Another example is the choice of a job or a career. Is it better to work for a large, stable company where job security is good but the chances of advancement are limited, or to join a new venture, which offers less job security but quicker advancement?

To answer these questions, we must be able to quantify risk so as to be able to compare the riskiness and alternative choices. We must see how people can deal with risk or reduce risk — by diversification, by buying insurance, etc. or by investing in additional information. In different situations, people must choose the amount of risk they wish to bear. To analyse risk quantitatively, we need to know all possible outcomes of a particular action and the likelihood that each outcome will occur.

## 2.2 UNCERTAINTY AND CHOICE UNDER UNCERTAINTY

Very often we have to select from a number of alternatives which differ in the risk the consumer has to bear. This is seen in cases like

insurance and gambling. When you take insurance policy (say for a fire in your house or car theft), you lose your premium (a small amount) to avoid the risk of losing your house or a car (a large value). However, this situation may or may not occur. The loss of say house is probable and therefore uncertain. But it is certain that you lose your premium when you have paid it. Here we prefer certainty of small loss to uncertainty of a large loss.

The risk refers to a situation when the outcome of a decision is uncertain but where the probability of each possible outcome is known or can be estimated. The greater the variability of possible outcome, the greater the risk involved in making the decision.

The uncertainty refers to the situation where there is more than one possible outcome of a decision but where the probability of occurrence of each particular outcome is not known or even cannot be estimated.

Many of the choices that people make involve considerable uncertainty.

Sometimes we need to choose between risky ventures.

For example, what should we do with our savings? Should we invest in something safe, such as a bank savings account, or something riskier but more lucrative, such as the stock markets? Another example is the choice of a job or a career.

Is it better to work for a large, stable company where job security is good but the chances of advancement are limited, or to join a new venture, which offers less job security but quicker advancement?

To answer these and such questions, we must be able to quantify risk so as to be able to compare the riskiness and alternative choices.

People deal with risk or reduce risk — by diversification, by buying insurance, etc. or by investing in additional information. In different situations, people must choose the amount of risk they wish to bear. For the quantitative analysis of risk we need to know all possible outcomes of a particular action and the likelihood that each outcome will occur. Following methods are used like

**Probability:**

Probability refers to the likelihood that an outcome will occur. Suppose the probability that the oil exploration project is successful might be 1/4, and the probability that it is unsuccessful 3/4. Probability could be objective and subjective. Objective probability relies on the frequency with which certain events have occurred. Suppose we know from our experience that, of the last 100 offshore oil explorations, 1/4 have succeeded and 3/4 have failed. Then the probability of success of 1/4 is objective because it is based on the frequency of similar experiences.

If we toss an unbiased coin, we would obtain two outcomes namely head and tail. If we toss a coin for quite good number of times there are

say 50% or ½ chances of getting head or 50% or 1/2 chance of getting tail. Here the sum of the probabilities of all possible outcomes would be equal to 1. In case of tossing a coin, it is ½ + ½= 1.

In another example, let us assume that a person, from the shares of a company, has got 50% dividend in 5% periods, 30% dividend in 60% periods and 10% dividend in 35 percent periods. Here the three rates of dividend, 50,30 and 10 are exhaustive. Thus, in this case, the probability of getting a dividend of 50% is 5% or1/20, that of getting dividend of 30% is 60% or 12/20 and the probability of getting a dividend of 10% is 35% or 7/20, here

$$1/20+ 12/20+ 7/20 = 1.$$

But what if there are no similar past experiences to help measure probability? In these cases, objective measures of probability cannot be obtained, and a more subjective measure is needed. Subjective probability is the perception that an outcome will occur and the perception is based on a person's judgment or experience, but not on the frequency of outcome observed in the past.

Whatever be the interpretation of probability, it is used to calculate two important measures that help us describe and compare risky choices. One measure tells us the expected value and the other variability of the possible outcomes.

**Expected Value:**
The expected value of an uncertain event is a weighted average of the values associated with all possible outcomes, with the probabilities of each outcome used as weights. The expected value measures the central tendency. In the above example, dividend is a variable-its three values are 50%, 30% and 10% and their probabilities are 1/20, 12/20 and 7/20 respectively. In this case the expected value of the dividend is (1/20×50 + 12/20×30+7/20×10) % or 24%.

In another example Suppose we are considering an investment proposal in an offshore oil company with two possible outcomes: success yields a payoff of £40 per share, while failure yields a payoff of £20 per share.

**The expected value in this case is given by:**
Expected Value = Pr (success) (£40/share) + Pr (failure) (£20/share)
= 1/4 (£40/share) + 3/4 (£20/share) = £25/share.

More generally, if there are two possible outcomes having pay offs $X_1$ and $X_2$, and the probabilities of each outcome are given by Pr and $Pr_2$, then the expected value E(X) is: $E(X) = Pr_1X_1 + Pr_1 X_2$ …….. (1)

**Variability:**
The variability or dispersion of a variable is the extent to which its values are dispersed or scattered. For example, if the first set of values of variable are 30,35,40,45, and 50 and second set of values of

variables are 5,10, 30, 50 and 70. It is clear that the variability of second set is greater than the first one. Significance of variability, small or large, is important and is different in different cases. Suppose, the first set of values is the runs of a particular cricketer in five different matches and second set of values is runs in five matches of second cricketer. Here the significance of smaller variability in the first case and higher variability in the second case is that the first player is more consistent performer than the second player.

Suppose we are choosing between two sales jobs that have the same expected income (£1,500). The first is based on commission. The second job is salaried. There are two equally likely incomes under the first job — £2,000 for a good sales effort and £1,000 for a moderate effort. The second job pays £ 1510 most of the time, but would pay £510 in severance pay if the business goes burst.

| *Procedure* | *Conventional designations* | *Error variance* |
|---|---|---|
| 1. Retest with same form on different occasions | Coefficient of stability | Temporal fluctuation |
| 2. Retest with parallel form on different occasion | Coefficient of stability and equivalence | Temporal fluctuation and item specification |
| 3. Retest with parallel form on same occasion | Coefficient of equivalence | Item specificity |
| 4. Split half (odd-even or other parallel splits) | Coefficient of internal consistency | Item specificity |
| 5. Kuder-Richardson (and other measures of inter-item consistency) | Coefficient of internal consistency | Item specificity and heterogeneity |



The two jobs have the same expected income because .5 (£2,000) + .5 (£1,000) = .99 (£1,510) + 0.1 (£510) = £1,500. But the variability of the possible payoffs is different for the two jobs. The variability can be analysed by a measure that presumes that large differences between actual payoffs and the expected payoff, called deviations,

Following Table gives the deviations of actual incomes from the expected income for the two sales jobs:

**Table No. 2.1**

| **Deviations from Expected Income £** | | | | |
|---|---|---|---|---|
| | **Outcome 1** | **Deviation** | **Outcome 2** | **Deviation** |
| Job 1 | 2,000 | 500 | 1,000 | 500 |
| Job 2 | 1,510 | 10 | 510 | 990 |

In the first job, the average deviation is £500:
Thus, Average Deviation = .5 (£500) + .5 (£500) = £500

**For the second job, the average deviation is calculated as:**
Average Deviation = .99 (£10) + .01 (£990) = £19.80

The first job is, thus, substantially more risky than the second as the average deviation of £500 is much greater than the average deviation

of £19.80 for the second job. The variability can be measured either by the variance which is the average of the squares of the deviations of the payoffs associated with each outcome from their expected value or by the standard deviation ($\sigma^2$) which is the square root of the variance.

**The average of the squared deviations under job 1 is given by:**
Variance ($\sigma^2$) = .5 (£2, 50,000) + .5 (£2, 50,000) = £2, 50,000

The standard deviation is equal to the square root of £2, 50,000 or £500.

Similarly, the average of the squared deviations under Job 2 is given by:

Variance ($\sigma^2$) = .99 (£100) + .01 (£9, 80,100) = £9,900.

The standard deviation (a) is the square root of £9,900 or £99.50. We use variance or standard deviation to measure risk, the second job is less risky than the first. Both the variance and the standard deviation of the incomes earned are lower. The concept of variance applies equally well when there are many outcomes rather than just two.

**Decision-making:**
Suppose we are choosing between the two sales jobs described above. What job should we take? If we dislike risk; we will take the second job. It offers the same expected return as the first but with less risk. Now suppose we add £100 to each of the payoffs in the first job, so that the expected payoff increases from £1,500 to £1,600.

**The jobs can then be described as:**
Job 1: Expected Income = £1,600 Variance = £2, 50,000
Job 2: Expected Income = £1,500 Variance = £ 9,900

Job 1 offers a higher expected income but is substantially riskier than job 2. Which job is preferred depends on us. If we are risk-lovers, we may opt for the higher expected income and higher variance, but a risk-averse person might opt for the second. We need to develop a consumer theory to see how people might decide between incomes that differ in both expected value and in riskiness.

**Choice under Uncertainty: Preference towards Risk:**
We use the above job example to describe how people might evaluate risky outcomes, but the principles apply equally well to other choices. Here we concentrate on consumer choices generally, and on the utility that consumers derive from choosing among risky alternatives.

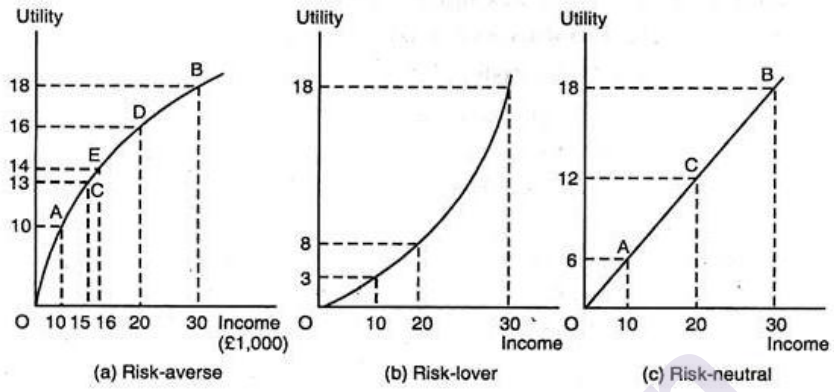
To simplify matters, we will consider the consumption of a single commodity, say, the consumer's income. We assume that consumers know probabilities and that payoffs are now measured in terms of utility rather than money.

Fig. 5.1(a) shows how we can describe one's preferences towards risk. The curve OB gives one's utility function, tells us the level of utility

that one can attain for each level of income. The level of utility increases from 10 to 16 to 18 as income increases from £10,000 to £20,000 to £30,000.

However, the marginal utility diminishes from 10 when income increases from 0 to £10,000, to 6 when income increases from £10,000 to £20,000, to 2 when income increases from £20,000 to £30,000.

**Table No. 2.1 Risk Aversion**




Now, suppose, we have an income of £15,000 and are considering a new but risky job that will either double our income to £30,000 or cause it to fall to £10,000. Each has a probability of 0.5. As Fig. 5.1(a) shows, the utility level associated with an income of £10,000 is 10 (point A), and the utility level associated with a level of £30,000 is 18 (point B). The risky job must be compared with the current job, for which utility is 13 (point C).

To evaluate the new job, we can calculate the expected value of the resulting income. Because we are measuring value in terms of utility, we must calculate the expected utility we can get. The expected utility is the sum of the utilities associated with all possible outcomes, weighed by the probability that each outcome will occur.

In this case, expected utility is E(U) = 1/2U (£10,000) + 1/2U (£30,000) = 0.5 (10) + 0.5 (18) = 14.

The new risky job is, thus, preferred to the old job because the expected utility of 14 is greater than the original utility of 13. The old job involved no risk — it guaranteed an income of £15,000 and a utility level of 13. The new job is risky, but it offers the prospect of both a higher expected income and a higher expected utility of 14. If we wished to increase our expected utility, we would take the risky job.

**Choice under Uncertainty: Different Preferences towards Risk:**
People differ in their willingness to bear risk. Some are risk-averse, some risk-lovers and some risk-neutral. A person who prefers a certain given income to a-risky job with the same expected income is known as risk-averse which is the most common attitude towards risk.

Most people not only insure against risks — such as, life insurance, health insurance, car insurance, etc. but also seek occupation with relatively stable wages.

Figure 2.1(a) applies to a person who is risk-averse. Suppose a person can have a certain income of £20,000 or a job yielding an income of £30,000 with probability 1/2 and an income of £10,000 with probability 1/2. As we have seen, the expected utility of the uncertain income is 14, an average of the utility at point A (10) and the utility at B (18), and is shown at E.

Now we can compare the expected utility associated with the risky job to the utility generated if £20,000 were earned without risk which is given by D (16) in Fig. 2.1(a). It is definitely greater than the expected utility with the risky job E (14).

A person who is risk-neutral is indifferent between earning a certain income and an uncertain income with the same expected income. In Fig. 2.1(c) the utility associated with a job generating an income between £10,000 and £30,000 with equal probability is 12, as is the utility of receiving a certain income of £20,000.

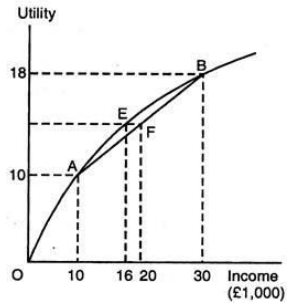
Fig. 2.1(b) shows the probability of risk-lover. In this case, the expected utility of an uncertain income that can be £10,000 with probability 1/2 or £30,000 with probability 1/2 is higher than the utility associated with a certain income of £20,000. As shown:
E(U) = 1/2U(£10,000) + 1/2V(£30,000) = 1/2(3) + 1/2(18) =10.5 > U(£20,000) = 8.

The main evidence of risk-loving is that people enjoy gambling. But very few people are risk-loving with respect to large amount of income or wealth. The risk premium is the amount that a risk-averse person would be willing to pay to avoid risk taking.

The magnitude of the risk premium depends on the risky alternatives that the person faces. The risk premium is determined in Fig. 5.2, which is the same utility function as in Fig. 2.1(a). An expected utility of 14 is achieved by a person who is going to take a risky job with an expected income of £20,000.

This is shown in Fig. 2.2 by drawing a horizontal line to the vertical axis from point F, which bisects the straight line AB. But the utility level of 14 can also be achieved if the person has a certain income of £16,000. Thus, the risk premium of £4,000, given by line EF, is the amount of income one would give up to leave him indifferent between the risky job and the safe one.

**Figure No. 2.2 Risk Premium**



How risk-averse a person is depends on the nature of the risk involved and on the person's income. Generally, risk-averse people prefer risks involving a smaller variability of outcomes. We saw that, when there are two outcomes, an income of £10,000 and £30,000 — the risk premium is £4,000.

We now consider a second risky job, involving a 0.5 probability of receiving an income of £40,000 and a utility level of 20 and a 0.5 probability of getting an income of 0. The expected value is also £20,000, but the expected utility is only 10.
Expected utility = .5U (£0) + .5U (£40,000) = 0 + .5(20) = 10.

Since the utility associated with having a certain income of £20,000 is 16, the person loses 6 units of utility if he is required to accept the job. The risk premium in this case is equal to £10,000 because the utility of a certain income of £10,000 is 10.

He can, thus, afford to give up £10,000 of his £20,000 expected income to have a certain income of £10,000 and will have the same level of expected utility. Thus, the greater the variability, the more a person is willing to pay to avoid the risky situation.

**Choice under Uncertainty: Reducing Risk:**
Sometimes consumers choose risky alternatives that suggest risk-loving rather than risk- averse behaviour, as the recent growth in state lotteries suggest. Nevertheless, in the face of a broad variety of risky situations, consumers are generally risk-averse. Now we describe three ways in which consumers can reduce risks diversification, insurance, and obtaining more information about choices and payoffs.

**Choice under Uncertainty: Diversification:**
Suppose that you are risk-averse and try to avoid risky situations as much as possible and you are planning to take a part-time selling job on a commission basis. You have a choice as to how to spend your time selling each appliance. Of course, you cannot be sure how hot or cold the weather will be next year. How should you apportion your time to minimize the risk involved in the sales job?

The risk can be minimized by diversification — by allocating time towards selling two or more products, rather than a single product. For example, suppose that there is a fifty-fifty chance that it will be a relatively hot year, and a fifty-fifty chance that it will be relatively cold.

**Gives the earnings you can make selling air-conditioners and heaters:**
**Table No. 2.2**

| **Income From Sale of Equipment** | | |
|---|---|---|
| | **Hot weather** | **Cold weather** |
| Air-conditioner sales | £ 30,000 | £ 12,000 |
| Heater sales | £ 12,000 | £ 30,000 |

If we decide to sell only air-conditioners or only heaters, our actual income will be either £12,000 or £30,000 and expected income will be £21,000 [.5(£30,000) + .5(£12,000)]. Suppose we diversify by dividing our time evenly between selling air-conditioners and heaters. .

Then our income will certainly be £21,000, whatever be the weather. If the weather is hot, we will earn £15,000 from air-conditioner sales and £6,000 from heater sales; if it is cold, we will earn £6,000 from air-conditioner sales and £ 15,000 from heater sales. In either case, by diversifying, we assure ourselves a certain income and eliminate all risks.

Diversification is not always easy. In our example, whenever the sales of one were strong, the sales of the other were weak. But the principle of diversification has a general application. As long as we can allocate our effort or investment funds towards a variety of activities, whose outcomes are not closely related, we can eliminate some risk.

**Choice under Uncertainty: Insurance:**

We have seen that risk-averse people will be willing to give up income to avoid risk. If, however, the cost of insurance is equal to the expected loss, risk-averse people will wish to buy enough insurance to offset losses they might suffer. The reasoning is implicit in our discussion of risk-aversion.

Buying insurance means a person will have the same income whether or not there is a loss, because the insurance cost is equal to the expected loss. For a risk-averse person, the guarantee of the same income, whatever be the outcome, generates more utility than would be the case if that person had a high income when there is no loss and a low income when a loss occurred.

Suppose a homeowner faces a 10% probability that his house will be burglarized and he will suffer a loss of £10,000. Let us assume that he has £50,000 worth of property.

**Table 2.3 shows his wealth with two possibilities — to insure or not to insure:**

**Table No. 2.3**

| Decision to Insure | | | |
|---|---|---|---|
| **Insurance** | **Burglary(Pr=.1)** | **No Burglary(Pr=.9)** | **Expected wealth** |
| No | £ 40,000 | £ 50,000 | £ 49,000 |
| Yes | £ 49,000 | £ 49,000 | £ 49,000 |

The decision to purchase insurance does not alter his expected wealth. It does smoothen it out over both possibilities. This generates a high level of expected utility to the house-owner, because the marginal utility in both situations is the same for the person who buys insurance. But when there is no insurance, the marginal utility in the event of a loss is higher than if no loss occurs. Thus, a transfer of wealth from the no-loss to the loss situation must increase total utility. And this transfer of wealth is exactly what is achieved through insurance.

Persons usually buy insurance from companies that specialise in selling it. Generally, insurance companies are profit-maximising firms that offer insurance because they know that, when they pool risk, they face very little risk.

This avoidance of risk is based on the law of large numbers, which tells us that although single events may be random and difficult to predict, the average outcome of many similar events may be predicted. For example, if one is selling automobile insurance, one cannot predict whether a particular driver will have an accident, but one can be reasonably sure, judging from past experience, about how many accidents a large group of drivers will have.

By operating on a large scale, insurance companies can be sure that the total premiums paid in will be equal to the total amount of money paid out. In our burglary example, a man knows that there is a 10% probability of his house being burgled; if it is, he will suffer a £10,000 loss. Prior to facing this risk, he calculated his expected loss of £1,000 (£10,000 x 0.1), but this is a substantial risk of loss.

Now suppose 100 people face this situation and all of them buy burglary insurance from a company. The insurance company charges each of them a premium of £1,000 which generates an insurance fund of £1, 00,000 from which losses can be paid.

The insurance company can rely on the law of large numbers which assures it that the expected loss for every individual is likely to be met. Thus, the total payout will be close to £1, 00,000 and the company need not worry about losing more than that amount.

Insurance companies are likely to charge premiums higher than the expected loss because they need to cover their administrative costs. Thus, many people may prefer to self-insurance rather than buy from an insure company. One way to avoid risk is to self-insure by diversifying.

**Choice under Uncertainty: Value of Information:**

The decision a consumer makes when outcomes are uncertain is based on limited information. If more information were available, the consumer could reduce risk. Since information is a valuable commodity, people will be prepared to pay for it. The value of complete information is the difference between the expected value with complete information and the expected value with incomplete information.

To see the value of information, suppose you are a manager of a store and must decide how many suits to order for the fall season. If you order 100 suits, your cost is £180 per suit, but if you order 50 suits, your cost would be £200. You know you will be selling for £300 each, but you are not sure what total sales would be.

All unsold suits could be returned but for half the price you paid for them. Without further information, you will act on the belief that there

is a 0.5 probability that 100 suits will be sold and a 0.5 probability that 50 will be sold.

**Table 2.4 gives the profit that you could earn in each of the two cases:**

**Table 2.4**

| Profits from Suits | | | |
|---|---|---|---|
| **Insurance** | **Burglary(Pr=.1 )** | **No Burglary(Pr=.9)** | **Expected wealth** |
| 1. Buy 50 suits | £ 5,000 | £ 5,000 | £ 5,000 |
| 2. Buy 100 suits | £ 1,500 | £ 12,000 | £ 6,750 |

Without more information, you would buy 100 suits if you were risk-neutral, taking the chance that your profit might be either £12,000 or £1,500. But if you were risk-averse, you might buy 50 suits for a guaranteed income of £5,000.

With complete information, you can make the correct suit order, whatever the sales might be. If sales were going to be 50 suits and you order for 50, you make a profit of £5,000. On the other hand, if sales were going to be 100 and you order for 100, you make a profit of £12,000. Since both outcomes are equally likely, your expected profit with complete information would be £8,500.

**The value of information is:**

| | |
|---|---|
| **Expected value with complete information** | **£8,500.00** |
| **– Expected value with uncertainty** | **– £6,750.00** |
| **Value of complete information** | **£1,750.00** |

Thus, it is worth paying up to £1,750.00 to obtain as accurate an information as possible.

**Choice under Uncertainty: Demand for Risky Assets:**

People are generally risk-averse. Given a choice, they prefer a fixed income to one that is as large on average that fluctuates randomly. Yet many of these people will invest all or part of their savings in stocks, bonds and other assets that carry some risk.

Why do risk-averse people invest in risky stocks either all or part of their investment? How do people decide how much risk to bear for the future? To answer these questions, we must examine the demand for risky assets.

**Choice under Uncertainty: Assets:**

An asset is something that provides a monetary flow to its owner. The monetary flow from owning an asset can take the form of an explicit payment, such as the rental income from an apartment building. Another explicit payment is the dividend on shares.

But sometimes the monetary flow from ownership of an asset is implicit; it takes the form of an increase or decrease in the price or value of the asset — a capital gain or a capital loss.

A risky asset provides a monetary flow that is in part random, which means, the monetary flow is not known with certainty in advance. A share of a company is an obvious example of a risky asset — one cannot know whether the price of the stock will rise or fall over time, and one cannot even be sure that the company will continue to pay the same dividend per share.

Although people often associate risk with the stock market, most other assets are also risky.

The corporate bonds are example of this — the corporation that issued the bonds could go bankrupt and fail to pay bond owners their returns. Even long-term government bonds that mature in 10 or 20 years are risky.

Although it is unlikely that government will go bankrupt, the rate of inflation could increase and make future interest payments and the eventual repayment of principal worth less in real terms, and, thus, reduce the value of the bonds.

In contrast to risky assets, we can call an asset riskless if it pays a monetary flow that is certain. Short-term government bonds — known as Treasury Bills — are risk-free assets because they mature within a short period, there is very little risk of an unexpected increase in inflation.

And one can also be confident that government will not default on the bond. Other examples of riskless assets include passbook savings accounts in banks and building societies or short- term certificate of deposit.

**Choice under Uncertainty: Asset Returns:**

People buy and hold assets because of the monetary flows they provide. Assets may be compared in terms of their monetary flow relative to the price of asset. The return on an asset is the total monetary flow it provides as a fraction of its value. For example, a bond worth £1,000 today that pays out £100 this year has a return of 10%.

When people invest their savings in stocks, bonds or other assets, they usually hope to earn a return that exceeds that rate of inflation, so that, by delaying consumption, they can consume more in the future. Thus, we often express the return on an asset in real terms which means return less the rate of inflation. For example, if the annual rate of inflation had been 5%, the bond would have yielded real return of 5%.

Since most assets are risky, an investor cannot know in advance what return they are going to yield in future. However, one can compare assets by looking at their expected returns which is just the expected value of its return. In a particular year, the actual return may be higher

or lower than expected, but over a long period the average return should be close to the expected return.

Different assets have different expected returns. Table 5.6 shows that the expected real return on Treasury Bills has been less than 1%, while the real return for a representative stock on the London Stock Market has been almost 9%.

Why would a person buy a Treasury bill when the expected return on stocks is so much higher? The answer is that the demand for an asset depends not only on expected return, but also on its risk.

One measure of risk, the standard deviation (σ) of the real return, is equal to 21.2% for common stock, but only 8.3% for corporate bonds, and 3.4% for Treasury Bills, as Table 5.6 shows. Clearly, the higher the expected return on investment, the greater the risk involved. As a result, a risk-averse investor must balance expected return against risk.

**Table No. 2.5**

| **Investment risk and Return** | | |
|---|---|---|
| **Insurance** | **Real Rate of Return (%)** | **Risk (Standard Deviation, σ, %)** |
| Common Stock | 8.8 | 21.2 |
| Long term corporate bonds | 2.1 | 8.3 |
| Treasury Bills | 0.4 | 3.4 |

**Choice under Uncertainty: Trade-Off between Risk and Return:**
Suppose a person has to invest his savings in two assets — riskless Treasury Bills, and a risky representative group of stocks. He has to decide how much of his savings to invest in each of these two assets. This is analogous to the consumer's problem of allocating a budget between two goods x and y.

Let us denote the risk-free return on the Treasury Bill by $R_f$, where the expected and actual returns are the same. Also, assume the expected return from investing in the stock market is $R_m$, and the actual return is $Y_m$.

The actual return is risky. At the time of investment decision, we know the likelihood of each possible outcome, but we do not know what particular outcome will occur. The risky asset will have a higher expected return than the risk-free asset ($R_m > R_f$) Otherwise, risk-averse investors would invest only in Treasury Bills and none at all in stocks.

To determine how much he will invest in each asset, let us assume b is the fraction of his savings placed in the stock market, and (1 – b) the fraction used to purchase Treasury Bills. The expected return on his total portfolio, $R_p$, is a weighted average of the expected return on the two assets

$$R_p = bR_m + (1 - b)R_f \text{..... (2)}$$

Suppose, the stock market's expected return is 12%. Treasury Bills pay 4%, and b = 1/2. Then $R_p$ = 8%. How risky is this portfolio? The riskiness can be measured by the variance of the portfolio's return. Let us assume the variance of the risky stock market investment is $\sigma^2_m$ and the standard deviation is $\sigma_m$. We can show that the σ of the portfolio is the fraction of the portfolio invested in the risky asset times the o of that asset:

$$\sigma_p = b\sigma_m \ldots\ldots\ldots (3)$$

**Choice under Uncertainty: Investor's Choice Problem:**

To determine how our investor should choose this fraction b, we must first show his risk- return trade-off analogous to the budget line of a consumer. To see this trade-off, we can rewrite equation (2) as

$$R_p = R_f + b(R_m - R_f).$$

From equation (3), we see that $b = \frac{\sigma_p}{\sigma_m}$, so that $R_p = R_f + \frac{(R_m - R_f)}{\sigma_m}\sigma_p$ .....(4)

The slope of the budget line is $R_m - R/\sigma_m$, which is the price of risk as shown in Fig. 2.3. Three indifference curves are drawn; each curve shows combinations of risk and return that have an investor equally satisfied. The curves are upward-sloping because a risk-averse investor will require a higher expected return if he is to bear a greater amount of risk. The utility-maximising investment portfolio is at the point where indifference curve $U_2$ is tangent to the budget line.

**Fig No. 2.3**

Choosing between risk and return

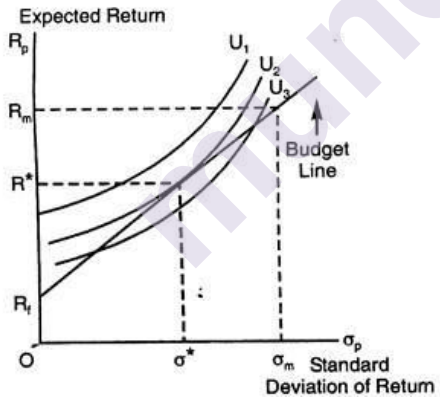
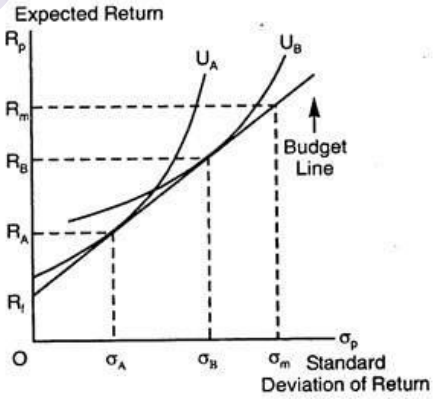



**Fig No. 2.4**

Choice of two different investors

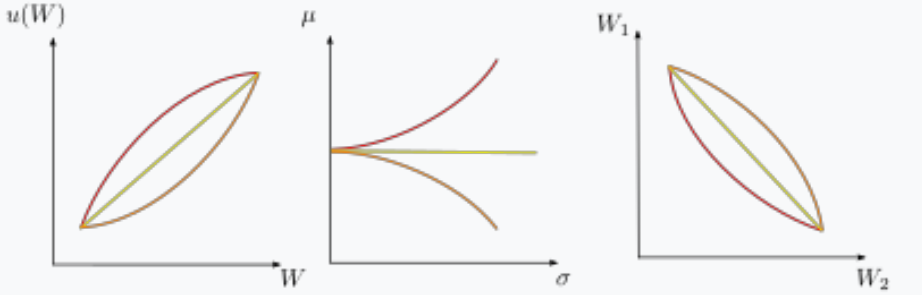



**Choice under Uncertainty :Two Different Attitudes to Risk: Choice under. Two Different Investors Choice with Different Attitudes to Risk:**

Investor A is risk-averse. His portfolio will consist mostly of the risk-free asset, so his expected return, $R_A$, will be only slightly greater than the risk-free return, but the risk $\sigma_A$ will be small. Investor B is less risk-averse. He will invest a large fraction of his funds in stocks. The expected return on his portfolio, $R_B$, will be larger, but the return will also be riskier.

## 2.3 MEASURES OF RISK AVERSION

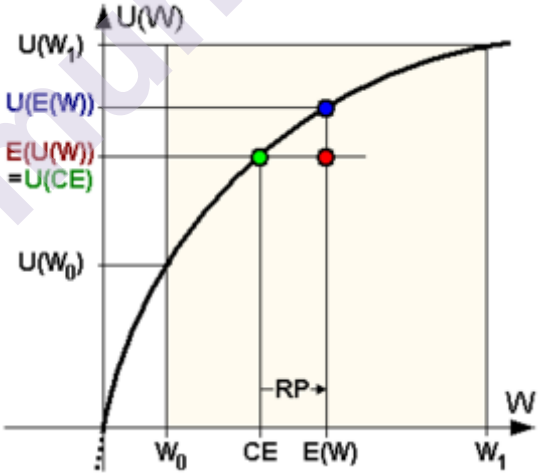
**Fig No. 2.5**




**In the above diagram, *Left graph*:** A risk averse utility function is concave (from below), while a risk loving utility function is convex. *Middle graph*: In standard deviation-expected value space, risk averse indifference curves are upward sloped. *Right graph*: With fixed probabilities of two alternative states 1 and 2, risk averse indifference curves over pairs of state-contingent outcomes are convex.

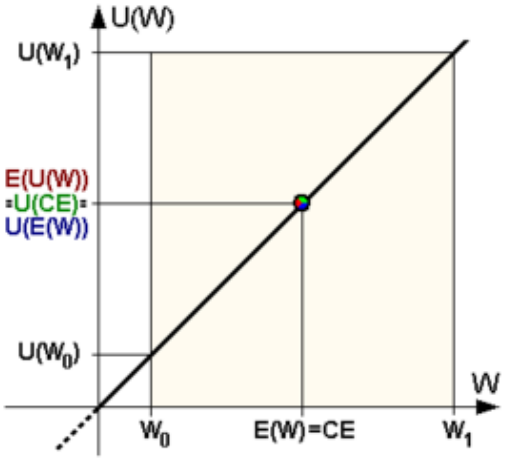
In economics and finance, **risk aversion** is the tendency of people to prefer outcomes with low uncertainty to those outcomes with high uncertainty, even if the average outcome of the latter is equal to or higher in monetary value than the more certain outcome. Risk aversion explains the inclination to agree to a situation with a more predictable, but possibly lower payoff, rather than another situation with a highly unpredictable, but possibly higher payoff. For example, a risk-averse investor might choose to put their money into a bank account with a low but guaranteed interest rate, rather than into a stock that may have high expected returns, but also involves a chance of losing value.

**Fig No. 2.6**



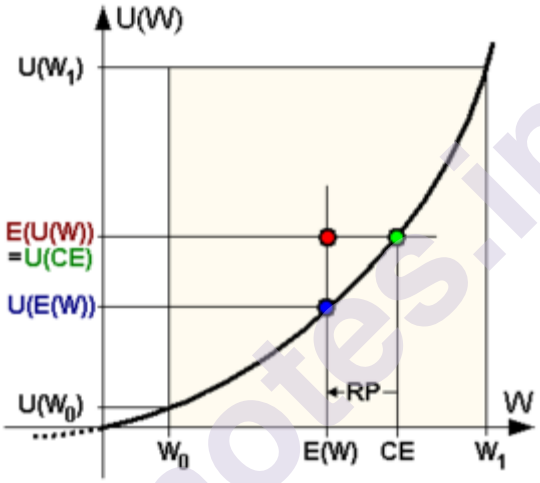

Utility function of a risk-averse (risk-avoiding) individual

**Fig No. 2.7**




Utility function of a risk-neutral individual

**Fig No. 2.8**




Utility function of a risk-loving (risk-seeking) individual

**CE – Certainty equivalent; E(U(W)) – Expected value** of the utility (expected utility) of the uncertain payment **W**; **E(W)** – Expected value of the uncertain payment; **U(CE) – Utility of the certainty equivalent; U(E(W))** – Utility of the expected value of the uncertain payment; **$U(W_0)$** – Utility of the minimal payment; **$U(W_1)$** – Utility of the maximal payment; **$W_0$** – Minimal payment; **$W_1$** – Maximal payment; **RP – Risk premium**

A person is given the choice between two scenarios: one with a guaranteed payoff, and one with a risky payoff with same average value. In the former scenario, the person receives $50. In the uncertain scenario, a coin is flipped to decide whether the person receives $100 or nothing. The expected payoff for both scenarios is $50, meaning that an individual who was insensitive to risk would not care whether they took the guaranteed payment or the gamble. However, individuals may have different **risk attitudes**.

**A person is said to be:**

- **risk averse** (or **risk avoiding):** if they would accept a certain payment (certainty equivalent) of less than $50 (for example, $40), rather than taking the gamble and possibly receiving nothing.

- **risk neutral:** if they are indifferent between the bet and a certain $50 payment.
- **risk loving** (or **risk seeking):** if they would accept the bet even when the guaranteed payment is more than $50 (for example, $60).

The average payoff of the gamble, known as its expected value, is $50. The smallest dollar amount that an individual would be indifferent to spending on a gamble or guarantee is called the **certainty equivalent**, which is also used as a measure of risk aversion. An individual that is risk averse has a certainty equivalent that is smaller than the prediction of uncertain gains. The risk premium is the difference between the expected value and the certainty equivalent. For risk-averse individuals, risk premium is positive, for risk-neutral persons it is zero, and for risk-loving individuals their risk premium is negative.

**Utility of money:**
In expected utility theory, an agent has a utility function $u(c)$ where $c$ represents the value that he might receive in money or goods (in the above example $c$ could be $0 or $40 or $100).

The utility function $u(c)$ is defined only up to positive affine transformation – in other words, a constant could be added to the value of $u(c)$ for all $c$, and/or $u(c)$ could be multiplied by a positive constant factor, without affecting the conclusions.

An agent possesses risk aversion if and only if the utility function is concave. For instance $u(0)$ could be 0, $u(100)$ might be 10, $u(40)$ might be 5, and for comparison $u(50)$ might be 6.

The expected utility of the above bet (with a 50% chance of receiving 100 and a 50% chance of receiving 0) is

$$E(u) = \big(u(0) + u(100)\big)/2$$

and if the person has the utility function with $u(0)=0$, $u(40)=5$, and $u(100)=10$ then the expected utility of the bet equals 5, which is the same as the known utility of the amount 40. Hence the certainty equivalent is 40.

The risk premium is ($50 minus $40) =$10, or in proportional terms

$$(\$50 - \$40)/\$40$$

or 25% (where $50 is the expected value of the risky bet: This risk premium means that the person would be willing to sacrifice as much as $10 in expected value in order to achieve perfect certainty about how much money will be received. In other words, the person would be indifferent between the bet and a guarantee of $40, and would prefer anything over $40 to the bet.

In the case of a wealthier individual, the risk of losing $100 would be less significant, and for such small amounts his utility function would be likely to be almost linear. For instance, if u(0) = 0 and u(100) = 10, then u(40) might be 4.02 and u(50) might be 5.01.

The utility function for perceived gains has two key properties: an upward slope, and concavity. (i) The upward slope implies that the person feels that more is better: a larger amount received yields greater utility, and for risky bets the person would prefer a bet which is first-order stochastically dominant over an alternative bet (that is, if the probability mass of the second bet is pushed to the right to form the first bet, then the first bet is preferred). (ii) The concavity of the utility function implies that the person is risk averse: a sure amount would always be preferred over a risky bet having the same expected value; moreover, for risky bets the person would prefer a bet which is a mean-preserving contraction of an alternative bet (that is, if some of the probability mass of the first bet is spread out without altering the mean to form the second bet, then the first bet is preferred).

**Measures of risk aversion under expected utility theory:**
There are multiple measures of the risk aversion expressed by a given utility function. Several functional forms often used for utility functions are expressed in terms of these measures.

**Absolute risk aversion:**
The higher the curvature of $(c)$ , the higher the risk aversion. However, since expected utility functions are not uniquely defined (are defined only up to affine transformations), a measure that stays constant with respect to these transformations is needed rather than just the second derivative of $(c)$ . One such measure is the **Arrow–Pratt measure of absolute risk aversion** (**ARA**), after the economists Kenneth Arrow and John W. Pratt, also known as the **coefficient of absolute risk aversion**, defined as

$$A(c) = -\frac{u''(c)}{u'(c)}$$

Where $u'(c)$ and $u''(c)$ denote the first and second derivatives with respect to c of $u(c)$. For example, if $u(c) = \alpha + \beta ln(c)$, so $u'(c) = \beta/c$ and $u''(c) = -\beta/c^2$, then $A(c) = 1/c.$ Note now $A(c)$ does not depend on $\alpha$ and $\beta$, so affine transformations of $u(c)$ do not change it.

**The following expressions relate to this term:**
Exponential utility of the form $u(c) = 1 - e^{-\alpha c}$ is unique in exhibiting *constant absolute risk aversion* (CARA): $A(c) = \alpha$ is constant with respect to *c*.

Hyperbolic absolute risk aversion (HARA) is the most general class of utility functions that are usually used in practice (specifically, CRRA (constant relative risk aversion, see below), CARA (constant absolute risk aversion), and quadratic utility all exhibit HARA and are often used because of their mathematical tractability). A utility function

exhibits HARA if its absolute risk aversion is a hyperbolic function, namely

$$A(c) = -\frac{u''(c)}{u'(c)} = \frac{1}{ac+b}$$

The solution to this differential equation (omitting additive and multiplicative constant terms, which do not affect the behavior implied by the utility function) is:

$$u(c) = \frac{(c-c_s)^{1-R}}{1-R},$$

where R=1/a and $c_8$= -b/a . Note that when a=0, this is CARA, as A(c) = 1/b = const, and when b=0, this is CRRA as cA(c) = 1/a=const .

*Decreasing/increasing absolute risk aversion* (DARA/IARA) is present if A(c)is decreasing/increasing. Using the above definition of ARA, the following inequality holds for DARA:

$$\frac{\partial A(c)}{\partial c} = -\frac{u'^{(c)}u'''(c) - \left[u''^{(c)}\right]^2}{[u'(c)]^2} < 0$$

and this can hold only if u'''(c) >0 . Therefore, DARA implies that the utility function is positively skewed; that is, u'''(c) >0 . Analogously, IARA can be derived with the opposite directions of inequalities, which permits but does not require a negatively skewed utility function (u'''(c)<0 ) An example of a DARA utility function is u(c) = log (c) , with A(c)=1/c, while u(c)= C-αc$^2$, with A(c)=2α/(1-2αc)would represent a quadratic utility function exhibiting IARA.

Experimental and empirical evidence is mostly consistent with decreasing absolute risk aversion.

Contrary to what several empirical studies have assumed, wealth is not a good proxy for risk aversion when studying risk sharing in a principal-agent setting. Although A(C)= $-\frac{u''(c)}{u'c}$ is monotonic in wealth under either DARA or IARA and constant in wealth under CARA, tests of contractual risk sharing relying on wealth as a proxy for absolute risk aversion are usually not identified.

**Relative risk aversion:**

The **Arrow–Pratt measure of relative risk aversion** (RRA) or **coefficient of relative risk aversion** is defined as

$$R(c) = cA(c) = \frac{-cu''}{u'(c)}$$

Unlike ARA whose units are in $\$^{-1}$, RRA is a dimension-less quantity, which allows it to be applied universally. Like for absolute risk aversion, the corresponding terms *constant relative risk aversion* (CRRA) and *decreasing/increasing relative risk aversion* (DRRA/IRRA) are used. This measure has the advantage that

it is still a valid measure of risk aversion, even if the utility function changes from risk averse to risk loving as $c$ varies, i.e. utility is not strictly convex/concave over all $c$. A constant RRA implies a decreasing ARA, but the reverse is not always true. As a specific example of constant relative risk aversion, the utility function $u(c) = \log(c)$ implies

RRA = 1. In intertemporal choice problems, the elasticity of intertemporal substitution often cannot be disentangled from the coefficient of relative risk aversion. The isoelastic utility function

$$u(c) = \frac{c^{1-p} - 1}{1 - p}$$

exhibits constant relative risk aversion with $R(c) = p$ and the elasticity of intertemporal substitution $\in u(c) = 1 / p$. When $p$=1,using l 'Hôpital's rule shows that this simplifies to the case of *log utility,* $u(c) = \log c$, and the income effect and substitution effect on saving exactly offset.

A time-varying relative risk aversion can be considered

**Implications of increasing/decreasing absolute and relative risk aversion:**

The most straightforward implications of increasing or decreasing absolute or relative risk aversion, and the ones that motivate a focus on these concepts, occur in the context of forming a portfolio with one risky asset and one risk-free asset. If the person experiences an increase in wealth, he/she will choose to increase (or keep unchanged, or decrease) the *number of dollars* of the risky asset held in the portfolio if *absolute* risk aversion is decreasing (or constant, or increasing). Thus economists avoid using utility functions such as the quadratic, which exhibit increasing absolute risk aversion, because they have an unrealistic behavioral implication.

Similarly, if the person experiences an increase in wealth, he/she will choose to increase (or keep unchanged, or decrease) the *fraction* of the portfolio held in the risky asset if *relative* risk aversion is decreasing (or constant, or increasing).

In one model in monetary economics, an increase in relative risk aversion increases the impact of households' money holdings on the overall economy. In other words, the more the relative risk aversion increases, the more money demand shocks will impact the economy.

**Portfolio theory:**

In **modern portfolio theory,** risk aversion is measured as the additional expected reward an investor requires to accept additional risk. If an investor is risk-averse, they will invest in multiple uncertain assets, but only when the predicted return on a portfolio that is uncertain is greater than the predicted return on one that is not uncertain will the investor will prefer the former. Here, the **risk-return spectrum** is relevant, as it results largely from this type of risk aversion. Here risk is measured as the standard deviation of the return

on investment, i.e. the square root of its variance. In advanced portfolio theory, different kinds of risk are taken into consideration. They are measured as the n-th root of the n-th central moment. The symbol used for risk aversion is A or $A_n$.

$$A = \frac{dE(c)}{d\sigma}$$

$$A_n = \frac{dE(c)}{d\sqrt[n]{\mu_n}}$$

**Limitations of expected utility treatment of risk aversion:**
Using expected utility theory's approach to risk aversion to analyze *small stakes decisions* has come under criticism. Matthew Rabin has showed that a risk-averse, expected-utility-maximizing individual who *from any initial wealth level [...] turns down gambles where she loses $100 or gains $110, each with 50% probability will turn down 50–50 bets of losing $1,000 or gaining any sum of money.*

Rabin criticizes this implication of expected utility theory on grounds of implausibility—individuals who are risk averse for small gambles due to diminishing marginal utility would exhibit extreme forms of risk aversion in risky decisions under larger stakes. One solution to the problem observed by Rabin is that proposed **by prospect theory and cumulative prospect theory**, where outcomes are considered relative to a reference point (usually the status quo), rather than considering only the final wealth.

Another limitation is the reflection effect, which demonstrates the reversing of risk aversion. This effect was first presented by Kahneman and Tversky as a part of the prospect theory, in the **behavioral economics** domain. The reflection effect is an identified pattern of opposite preferences between negative as opposed to positive prospects: people tend to avoid risk when the gamble is between gains, and to seek risks when the gamble is between losses. For example, most people prefer a certain gain of 3,000 to an 80% chance of a gain of 4,000. When posed the same problem, but for losses, most people prefer an 80% chance of a loss of 4,000 to a certain loss of 3,000.

The reflection effect (as well as the **certainty effec**t) is inconsistent with the expected utility hypothesis. It is assumed that the psychological principle which stands behind this kind of behavior is the overweighting of certainty. Options which are perceived as certain are over-weighted relative to uncertain options. This pattern is an indication of risk-seeking behavior in negative prospects and eliminates other explanations for the certainty effect such as aversion for uncertainty or variability.

The initial findings regarding the reflection effect faced criticism regarding its validity, as it was claimed that there are insufficient evidence to support the effect on the individual level. Subsequently, an extensive investigation revealed its possible limitations, suggesting that

the effect is most prevalent when either small or large amounts and extreme probabilities are involved.

**Public understanding and risk in social activities:**
In the real world, many government agencies, e.g. **Health and Safety Executive**, are fundamentally risk-averse in their mandate. This often means that they demand (with the power of legal enforcement) that risks be minimized, even at the cost of losing the utility of the risky activity. It is important to consider the **opportunity cost** when mitigating a risk; the cost of not taking the risky action. Writing laws focused on the risk without the balance of the utility may misrepresent society's goals. The public understanding of risk, which influences political decisions, is an area which has recently been recognised as deserving focus. In 2007 Cambridge University initiated the Winton Professorship of the Public Understanding of Risk, a role described as outreach rather than traditional academic research by the holder, David Spiegelhalter.

**Children:**
Children's services such as schools and playgrounds have become the focus of much risk-averse planning, meaning that children are often prevented from benefiting from activities that they would otherwise have had. Many playgrounds have been fitted with impact-absorbing matting surfaces. However, these are only designed to save children from death in the case of direct falls on their heads and do not achieve their main goals. They are expensive, meaning that less resources are available to benefit users in other ways (such as building a playground closer to the child's home, reducing the risk of a road traffic accident on the way to it), and—some argue—children may attempt more dangerous acts, with confidence in the artificial surface. Shiela Sage, an early years school advisor, observes "Children who are only ever kept in very safe places, are not the ones who are able to solve problems for themselves. Children need to have a certain amount of risk taking ... so they'll know how to get out of situations."

**Game shows and investments:**
One experimental study with student-subject playing the game of the TV show **Deal or No Deal** finds that people are more risk averse in the limelight than in the anonymity of a typical behavioral laboratory. In the laboratory treatments, subjects made decisions in a standard, computerized laboratory setting as typically employed in behavioral experiments. In the limelight treatments, subjects made their choices in a simulated game show environment, which included a live audience, a game show host, and video cameras. In line with this, studies on investor behavior find that investors trade more and more speculatively after switching from phone-based to online trading and that investors tend to keep their core investments with traditional brokers and use a small fraction of their wealth to speculate online.

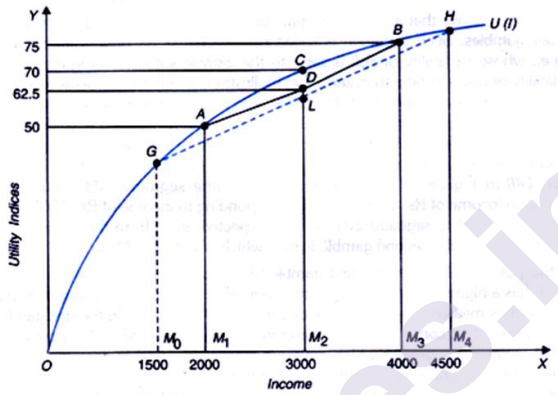
**Risk Aversion:**
People differ greatly in their attitudes towards risks. In Bernoullis hypothesis a person whose marginal utility of money declines will refuse to accept a fair gamble. A fair game or gamble is one in which

the expected vale of income from a gambler = same amount of income with certainty. The person who refuses a fair gamble is a risk averse. Thus, risk averter is one who prefers a given income with certainty to a risky gamble with the same expected value of income. Risk aversion is the most common attitude towards risk. It is because of the attitude of risk aversion that people insure against various kinds of risks such as burning of house, illness etc.

This attitude of risk aversion can be explained with N-M method of measuring expected utility. Marginal utility of income of risk averter diminishes as his income increases.

**Fig No. 2.9**

**The Neumann-Morgenstem Concave Utility Curve of a Risk-Averter**




U(I) is the N-M utility function curve. It starts from the origin and has a positive slope i.e. individual prefers more income to less. The curve is concave to the origin showing that the marginal utility of income of a person decreases as his income increases. Therefore, the utility curve represents the case of risk-averter or the attitude of risk-aversion. For example, with income of Rs. 2000/-, the persons utility is 50 which rises to 70 when his income increases to Rs. 3000/-. As income rises to Rs.4000/-, utility rises to 75.

Now, suppose the person's current income is Rs. 3000/-. He is offered a fair gamble in which he has a 50-50 chance of winning or loosing Rs. 1000/- . Thus , the probability of winning is ½ or 0.5. If he wins the game, his income will rise to Rs. 4000/- and if he looses the gamble, his income will fall to Rs. 2000/-. The expected money value of his income in this situation of uncertain outcome is given by: E(V) = 1/2 × 4000 +1/2 × 2000 = Rs.3000/-. If he rejects the gamble he will have the present income (i.e. Rs. 3000) with certainty. Though the expected value of his uncertain income prospect = his income with certainty, a risk averter will not accept his gamble. This is because as he acts on the basis of expected utility of his income in the uncertain situation (i.e. Rs. 4000/- if he wins and Rs.2000/- if he loses) can be obtained as Expected Utility(EU) = $\pi$ U (Rs. 4000 + 1− $\pi$ U (Rs. 2000). The diagram shows the utility of a person from Rs. 4000 is 75 (Point B) and utility from Rs. 2000 is 50 ( Point A), the expected utility from this uncertain prospect will be

$$E(U) = 1/2\ (75) + 1/2\ (50)$$
$$= 37.5 + 25 = 62.5$$

In N-M utility curve U(I) the expected utility can be found by joining points A (corresponding to Rs.2000) and point B (corresponding to Rs. 4000)by a straight-line segment AB---- then reading a point on it corresponding to the expected value of the gamble Rs. 3000, the expected value of the utility is M_(2 )D (=62.5) < M_(2 )C or Rs.70 which is the utility of income of Rs.3000 with certainty. Therefore, a person will not gamble. His rejection of gamble is due to the diminishing marginal utility of money income for him. The gain in utility from Rs.1000in case he wins < the loss in utility from Rs. 1000 if he loses the gamble. Therefore, the expected utility from the uncertain income prospect is less than the utility he obtains from the same income with certainty.

In case marginal utility of money income decreases a person will avoid fair gambles. Such a person is called risk averter as he prefers an income with certainty (i.e. whose variability or risk is zero) to the gamble with the same expected value (where variability or risk is > 0). For example, person with a certain income (Y) of Rs. 3000, two fair gambles are offered to him. First, a 50:50 chance of winning or losing Rs.1000, as before and second a 50:50 chance of winning or losing Rs.1500. With the even chance of winning or losing the expected value of income in the second gamble will be

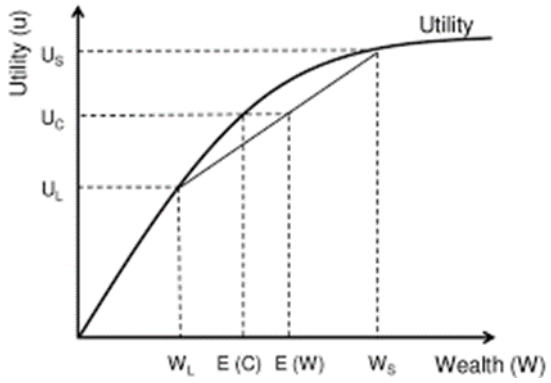
1/2 (1500) +1/2 (4500) = Rs. 3000. On N-M curve we draw a straight-line segment GH by joining (G---Y Rs. 1500 and H----Y Rs. 4500). It shows expected utility from the expected money value of Rs. 3000 from the second gamble is M_(2 )L < M_(2 )D of first gamble. The person will prefer the first gamble which has low variability to the second gamble which has higher degree of variability of outcome.

Risk Aversion and Insurance

A risk averse person is always ready to make some payments in order to avoid the risk-facing him. It means the risk averter would like to pay money to someone, say an insurance company, if he is assured a given income with certainty = the expected to value of the gamble with uncertain outcome.

For example, suppose a person has an income of Rs. 1,00,000 from the house he owns. The risk he faces is that if the house is burnt down by a fire he will suffer a loss of Rs. 40,000 in his income. Let us further suppose that the probability of his house being burnt down is 0.5. This is explained in the following diagram

**Fig No. 2.10**



If his house catches fire and burns down, his income will be reduced to Rs 60,000 (W_L). The expected value of the uncertain prospect is ½ (1,00,000 ( W_S)) + ½ (60,000) = Rs, 80,000 (E(W)). A straight lie segment is drawn between the utility points of the two uncertain outcomes of Rs. 1,00,000 and Rs. 60,000. Utility of the expected value of Rs. 80,000 is M_(2)D i.e. line drawn from E(w) or 60. When we draw the line from U_c, we find that a certain income= income 70000 i.e. E(c) also yields the same utility as the expected value of the gamble (Rs. 80,000). This means that a risk averse person will be willing to pay premium to the insurance company up to the maximum of Rs. 30,000 (Rs. 1,00,000 – Rs. 70,000) provides the insurance company agrees the restore his loss of Rs. 40,000 in case his house catches fire and burns down. Going in for insurance guarantees a person to have the sure income whether or not there is loss due to fire. Since risk aversion is the most common attitude, many people buy enough insurance against various types of risks.

**Fig No. 2.11**

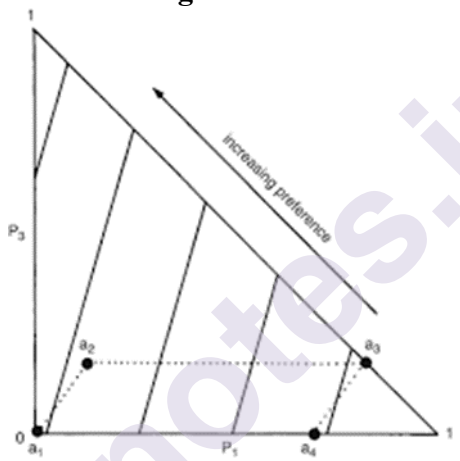
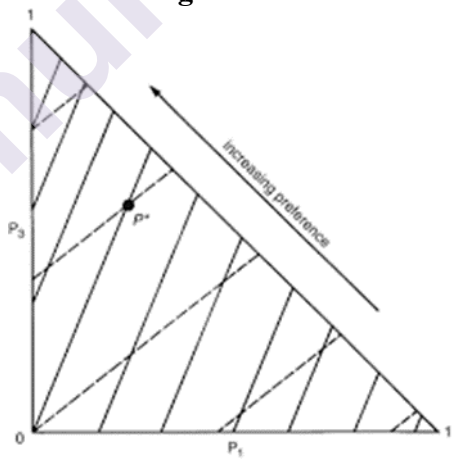



**Fig No. 2.11**




**Recent Analysis of the Speculative Behaviour:**

Let us now discuss the behaviour of economic agents such as investors and producers under the conditions of risk and uncertainty. This relates to the behaviour in the context of demand for money or liquidity. Here, let us take an account of the contributions of Prof. Baumol and James Tobin. It will help us to examine how the investors avert or minimise risk.

**Baumol's Inventory Demand Analysis:**
Transactions Demand: J.M Keynes, in his General Theory analysed total demand for money under three categories to satisfy transactions, precautionary and speculative motive. Baumol mainly explained the transactions motive. It is stated that the demand for money is a function of the income and is directly proportional to the size of the income of the investor. According to Baumol the transactions demand for money, though directly proportional to the size of the income, is less than proportionate to the increments in the income. For this Baumol made use of inventory demand approach.

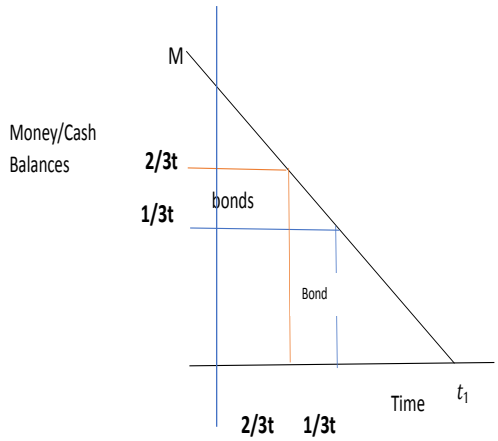
**Inventory Demand Method:** On account of transactions demand investors are required to possess large amount of money in cash or liquid form. This can be utilised for the payment of labour, raw material and such other charges. On this account, liquid money assets become necessary because of different timing of earning income and its expenditure. However, though holding money is a matter of convenience, it brings no income and hence/thus it is expensive. Moreover, large amount of money required to cover the whole volume of transactions over the entire income earning period is not immediately required or needed.

The next alternative open to the investors is that of investing part of their liquid assets in short term bonds or securities and earn interest income over such investment. Now the question is what proportions of investment in bonds and withdrawals in cash will be of optimum size from time to time. For this let us take a numerical example.

Let the total demand for liquid money at the beginning of the period t_0is Rs. 1,200which is to be carried over till the end of the period or beginning of the next period t_1. In other word the time interval t_0 to t_1let be of 12 months. Then the investor does not need the entire amount of Rs. 1200 at t_0 point of time. He can distribute it conveniently in the three instalments of Rs. 400 each. Therefore, he possesses Rs. 400 at t_0 period, instalments Rs.400 for 1/3t for four months and invests remaining Rs. 400 for2/3t or eight months. So that through out the year he has adequate cash resources and yet receives interest over Rs. 400 for four months and over Rs. 400 over eight months. The average demand for cash balances or liquid money sources is ½ M where M is the total annual demand of Rs. 1200. This is clear since average of the two months is

$$\frac{\frac{2}{3t} + 1/3t}{2} = \frac{1}{2}\,M$$

**Fig No. 2.12**




In the above diagram on the x-axis, we take time or of income earning period and on the Y axis we take total demand for money OM. Part of the OM, which is1/3 OM, is invested in bonds for 1/3t or four months and another 1/3OM invested in bonds for 2/3t or eight months. Now the next question is what is optimum proportion of investment in bonds and holding in cash balance?

**Variations in Investments:** The optimum level of demand for cash balances is flexible and will depend upon two types of factors.

On the one hand demand for money will depend upon b the value of brokerage charge, administrative expenses etc., of holding bonds (directly related), and

On the rate of interest or the income to be earned from the investment in bonds(inversely related).

This can be explained in the form of a function or an equation. Let C be the total inventory cost of making transactions, b the brokerage charges percentage, r the rate of interest or return on bonds, M/2 is the average demand for money, Y is the income of the firm and M the amount of withdrawal from the bonds and hence Y/M is the number of withdrawals. Then we have

$$c = r\frac{M}{2} + b\frac{Y}{M}$$

We want to find out optimum value of M which will minimise the value of C. This can be derived by differentiating C with respect to M and then setting it equal to zero.

$$\frac{dC}{dM} = \frac{d(r\frac{m}{2})}{dM} + \frac{d(b\frac{Y}{M})}{dM}$$

$= \frac{r}{2} - \frac{bY}{M^2}$, setting it equal to zero

$= \frac{r}{2} - - = 0$, or

$M^2 = \frac{2bY}{r}$, or

$M = \frac{\sqrt{2bY}}{r}$,

This is called square root formula of determining demand for money. This explains that (i) M the demand for money depends directly upon b the brokerage charges. Higher the rate of b lower will be the demand for bonds and more will be the demand for liquid money assets. (ii) Demand for M depends inversely on the rate of interest such that higher the interest rate more will be the investment in bonds and less will be the demand for money in cash balances.

**Diminishing Proportion:** Finally, Baumol has related variations in the demand for transaction cash balances and changes in the income Y. As the size of the income Y of the investor goes on increasing, he will no doubt require more and more cash balances. But he will also be able to invest more in the bonds. With large amounts of investment in the value of brokerage cost b will diminish. Therefore, relatively greater proportion of investment will be attractive, hence the proportion of cash balance demand will be restricted. Thus, he states that the transactions demand for money is no doubt the function of income but it progressively goes on falling in its proportion. For example, if transactions demand with Rs. 10,000 income is 10 percent (Rs. 1000) then with Rs. 50,000 income may be 8 percent (Rs. 4000) and with Rs. 80000 income may be 6 percent (Rs. 4800). Thus, with growing size of the income of the investor transactions demand also increases absolutely but diminishes relatively.

**James Tobin and Speculative Demand:**
Liquidity Preference: Keynes liquidity preference or speculative demand for money is based on certain assumptions. It depends on

Expectations about future rates of interest are inelastic.

Speculators or individuals like to hold either in cash money assets or bonds.

James Tobin in his 'Liquidity Preference as Behaviour towards Risk' has attempted fresh piece of analysis by removing these limitations of Keynesian theory

Instead of taking help of elasticity of expectations his theory is based on the assumption that expected value of gains or losses from holding interest basing assets is always equal to zero

It further assumes that a speculator or investor distributes his assets in both cash and bonds and not in either of them.

**B. Probability under Risk:**
In holding cash resources or money no risk is involved. But it brings no return in the form of interest income. To invest in bonds on the other hand is attractive activity which brings income but it also involves risk of capital gains or losses. More of such risk goes on increasing as the amount invested increases. The investors are prepared to accept such risk only when they are hopeful of adequate returns. If 'g' is expected gain or loss from the investment in bonds then investor will estimate the probability or act accordingly. The probability

distribution has zero expected value irrespective of the current rate of interest r on his investment. If M and B are the proportions of distribution of his assets between money and bonds then the total return R on his investment will have the value,
R=B (r+g) where O<B<1.

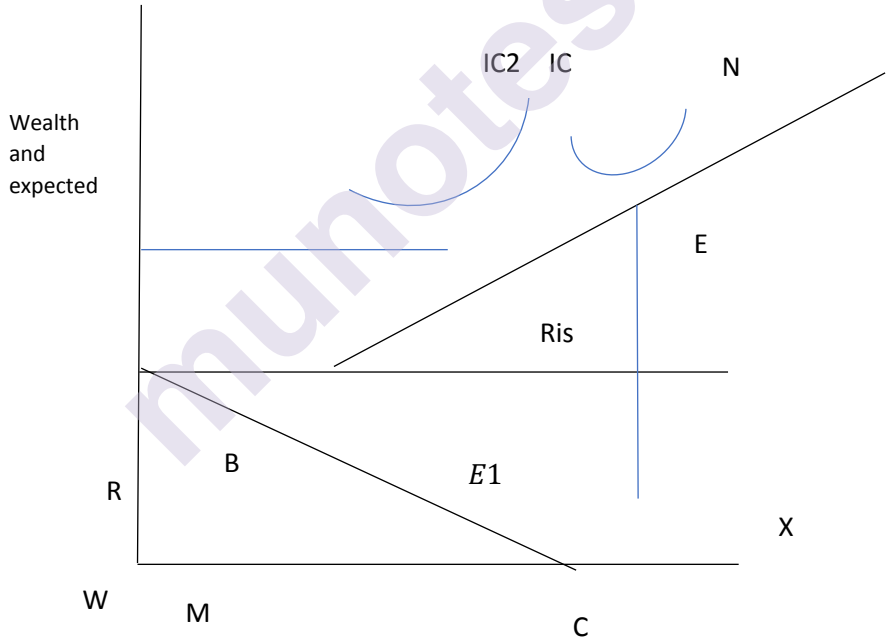
**Tobin then classifies investors into three categories:**
The risk lovers who desire to invest all their wealth on bonds and to maximise their risk. They are gamblers.

The there are plungers who will either invest all their wealth in bonds or will possess all in cash

Finally, there are risk averters or avoiders. They try to avoid risk associated with holding bonds rather than cash or money.

**Risk Avoidance:** It is worth to analyse behaviour of the third category of investors who are risk avoiders. Some people relate the amount of risk involved to the expected returns on the investment. Therefore, they try to distribute their assets both in bonds investment and money holding.

**Figure No. 2.13**

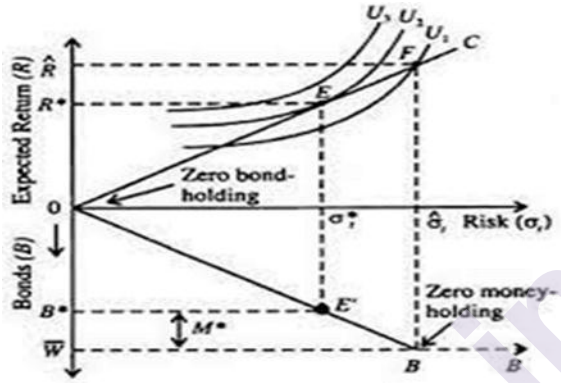



Risk on the X-axis and Returns on the Y-axis. $IC_1$ $and$ $IC_2$are upward sloping indifference curves of risk-bearing investors. It shows that investor expects higher and higher returns to undertake more and more risk. Wealth is measured along OW. Now ON is the budget line of investor and OC is downward counterpart showing proportional distribution of assets or wealth between bonds and money in accordance with the degree of risk. The investor is in equilibrium at point E i.e., tangency between$IC_1$ and budget line ON. A vertical line is drawn from E to meet OC at point $E_1$then optimum distribution of wealth between bonds and money can be determined. At point $E_1$the investor prefers to invest OR in bonds and holds remaining amount of

RW in money, thus risk is avoided by diversifying total wealth partly in bonds and partly in money.

The amount of risk involved is related by them to the expected returns on the investment. They will therefore appropriately try to distribute their assets both in bonds investment and money holding. It is then interesting to find out their preferences between risk and expected returns.

**Figure No. 2.14**
**Determination of the optimal portfolio**




Risk is shown on the X axis and Return along the Y axis. The expected return on the portfolio is the interest that can be earned on bonds. This depends on two things: (i) the interest rate and (ii) the proportion of the portfolio held in bonds. The total risk to which an individual is exposed depends on (i) the uncertainty concerning bond prices — that is, the uncertainty concerning future movements in market rate of interest, and (ii) the proportion of the portfolio held in bonds. Let us denote the expected total return by R and the total risk of the portfolio as a σt. If an individual holds all his wealth (W) in money and none in bonds, i.e., W = M + 0, both R and σt will be zero, as shown by the origin (point 0). With an increase in the proportion of bonds, i.e., W = M + B; as M falls and B increases, R and a, will both rise.

The opportunity line C is a locus of points showing the terms on which the individual investor can increase R at the cost of increasing σt. A movement along C from left to right shows that the investor increases his bond holding only by reducing his money holding.

The lower quadrant alternative portfolio allocations, resulting in different combinations of R and σt. The vertical axis measures bond holding. The amount of bonds (B) held in W increases as the investor moves down the vertical axis to a maximum of W.

The difference between W and B is the asset demand for money (M). The line OB in the lower part of the diagram shows the relationship between a, and B. As the proportion of B in W increases, σt also increases. This means that as the proportion of bonds in the portfolio increases, the total risk of the portfolio increases, too.

## 2.4 SUMMARY

People differ greatly in their attitudes towards risks. In Bernoullis hypothesis a person whose marginal utility of money declines will refuse to accept a fair gamble. A fair game or gamble is one in which the expected vale of income from a gambler = same amount of income with certainty. The person who refuses a fair gamble is a risk averse. Thus, risk averter is one who prefers a given income with certainty to a risky gamble with the same expected value of income. Risk aversion is the most common attitude towards risk. It is because of the attitude of risk aversion that people insure against various kinds of risks such as burning of house, illness etc.

## 2.5 QUESTIONS

Q1. Write a note on uncertainty and Choice under Uncertainty.

Q2. Explain the choice under uncertainty

Q3. Write a note on measures of Risk aversion

## 2.6 REFERENCES


- Gravelle H. and Rees R .(2004) : Microeconomics., 3rd Edition, Pearson Edition Ltd, New Delhi.
- Varian H (2000): Intermediate Microeconomics: A Modern Approach, 8th Edition, W.W.Norton and Company
- Gibbons R. A Primer in Game Theory, Harvester-Wheatsheaf, 1992
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.


*****

# MODULE II

# 3

# OLIGOPOLY MODELS- I

**Unit Structure**



## 3.0 OBJECTIVES

- To explore the knowledge of Oligopoly market and its Various Models
- To understand price and output determinations under Oligopoly Market.
- To know different types of equilibrium under oligopoly Model.
- To find out of variation in the equilibrium of Oligopoly Theories.

## 3.1 THE OLIGOPOLY MARKET: EXAMPLE, TYPES AND FEATURES

The **Oligopoly Market** characterized by few sellers, selling the homogeneous or differentiated products. In other words, the Oligopoly market structure lies between the pure monopoly and monopolistic competition, where few sellers dominate the market and have control over the price of the product.

**Under the Oligopoly market, a firm either produces:**

**Homogeneous product:** The firms producing the homogeneous products are called as Pure or Perfect Oligopoly. It is found in the producers of industrial products such as aluminium, copper, steel, zinc, iron, etc.

**Heterogeneous Product:** The firms producing the heterogeneous products are called as Imperfect or Differentiated Oligopoly. Such type of Oligopoly is found in the producers of consumer goods such as automobiles, soaps, detergents, television, refrigerators, etc.

The term oligopoly is derived from two Greek words: 'Oligos' means few and 'polis' means to sellers. Oligopoly is a market structure in which there are only a few sellers (but more than two) of the homogeneous or differentiated products. So, oligopoly lies in between monopolistic competition and monopoly.Oligopoly refers to a market situation in which there are a few firms selling homogeneous or differentiated products. Oligopoly is, sometimes, also known as 'competition among the few' as there are few sellers in the market and every seller influences and is influenced by the behaviour of other firms.

### 3.1.1 Types of Oligopoly:

**1. Pure or Perfect Oligopoly**:
If the firms produce homogeneous products, then it is called pure or perfect oligopoly. Though, it is rare to find pure oligopoly situation, yet, cement, steel, aluminium and chemicals producing industries approach pure oligopoly.

**2. Imperfect or Differentiated Oligopoly:**
If the firms produce differentiated products, then it is called differentiated or imperfect oligopoly. For example, passenger cars, cigarettes or soft drinks. The goods produced by different firms have their own distinguishing characteristics, yet all of them are close substitutes of each other.

**3. Collusive Oligopoly**:
If the firms cooperate with each other in determining price or output or both, it is called collusive oligopoly or cooperative oligopoly.

**4. Non-collusive Oligopoly:**
If firms in an oligopoly market compete with each other, it is called a non-collusive or non-cooperative oligopoly.

### 3.1.2 Features of Oligopoly:
**The main features of oligopoly are elaborated as follows:**

**1. Few firms:**
Under oligopoly, there are few large firms. The exact number of firms is not defined. Each firm produces a significant portion of the total output. There exists severe competition among different firms and each firm try to manipulate both prices and volume of production to outsmart each other. For example, the market for automobiles in India is an oligopolist structure as there are only few producers of automobiles.

The number of the firms is so small that an action by any one firm is likely to affect the rival firms. So, every firm keeps a close watch on the activities of rival firms.

**2. Interdependence**:
Firms under oligopoly are interdependent. Interdependence means that actions of one firm affect the actions of other firms. A firm considers the action and reaction of the rival firms while determining its price and output levels. A change in output or price by one firm evokes reaction from other firms operating in the market.

For example, market for cars in India is dominated by few firms (Maruti, Tata, Hyundai, Ford, Honda, etc.). A change by any one firm (say, Tata) in any of its vehicle (say, Indica) will induce other firms (say, Maruti, Hyundai, etc.) to make changes in their respective vehicles.

**3. Non-Price Competition:**
Under oligopoly, firms are in a position to influence the prices. However, they try to avoid price competition for the fear of price war. They follow the policy of price rigidity. Price rigidity refers to a situation in which price tends to stay fixed irrespective of changes in demand and supply conditions. Firms use other methods like advertising, better services to customers, etc. to compete with each other.

If a firm tries to reduce the price, the rivals will also react by reducing their prices. However, if it tries to raise the price, other firms might not do so. It will lead to loss of customers for the firm, which intended to raise the price. So, firms prefer non- price competition instead of price competition.

**4. Barriers to Entry of Firms:**
The main reason for few firms under oligopoly is the barriers, which prevent entry of new firms into the industry. Patents, requirement of large capital, control over crucial raw materials, etc, are some of the reasons, which prevent new firms from entering into industry. Only those firms enter into the industry which is able to cross these barriers. As a result, firms can earn abnormal profits in the long run.

**5. Selling Costs**:
Due to severe competition ‘and interdependence of the firms, various sales promotion techniques are used to promote sales of the product. Advertisement is in full swing under oligopoly, and many a times advertisement can become a matter of life-and-death. A firm under oligopoly relies more on non-price competition. Selling costs are more important under oligopoly than under monopolistic competition.

**6. Group Behaviour:**
Under oligopoly, there is complete interdependence among different firms. So, price and output decisions of a particular firm directly influence the competing firms. Instead of independent price and output strategy, oligopoly firms prefer group decisions that will protect the interest of all the firms. Group Behaviour means that firms tend to behave as if they

were a single firm even though individually they retain their independence.

**7. Nature of the Product:**
The firms under oligopoly may produce homogeneous or differentiated product.

i. If the firms produce a homogeneous product, like cement or steel, the industry is called a pure or perfect oligopoly.

ii. If the firms produce a differentiated product, like automobiles, the industry is called differentiated or imperfect oligopoly.

**8. Indeterminate Demand Curve:**
Under oligopoly, the exact behaviour pattern of a producer cannot be determined with certainty. So, demand curve faced by an oligopolist is indeterminate (uncertain). As firms are inter-dependent, a firm cannot ignore the reaction of the rival firms. Any change in price by one firm may lead to change in prices by the competing firms. So, demand curve keeps on shifting and it is not definite, rather it is indeterminate.

## 3.2 COURNOT'S DUOPOLY MODEL

A model of oligopoly was 1st put forward by Cournot a French economist in 1838. Cournot's model of oligopoly is one of the oldest theories of the behaviour of the individual firm and relate to non-collusive oligopoly. In the Cournot model it is assumed that an oligopolist thinks that his rival will keep their output fixed regardless of what he might do.

Another important model of non-collusive oligopoly was put forward by E. H .Chamberlin in his famous work "The theory of Monopolistic Competition". Chamberlin made an important improvement over the classical models of oligopoly, including that of Cournot. In sharp contrast to Cournot Chamberlin recognised is his model that oligopoly firms recognise their inter-dependence while fixing their output and price.

Cournot's Duopoly Model Augustine Cournot, a French economist, published his theory of duopoly in 1838. But it remained mainly unnoticed till 1880 when Walras called the attention of the economists to Cournot's work.

### 3.2.1 Assumptions:

1) Cournot takes the case of two identical mineral springs operated by two owners who are selling the mineral water in the same market. Their waters are identical. Therefore, his model relates to the duopoly with homogeneous products.
2) It is assumed by Cournot for the sake of simplicity that the owners operate mineral springs and sell water without incurring any cost of production.

3) The duopolists completely know the market demand for mineral water.
4) Cournot assumes that each duopolist believes that regardless of his actions and their effect on market price of the product, the rival firm will keep its output constant. .

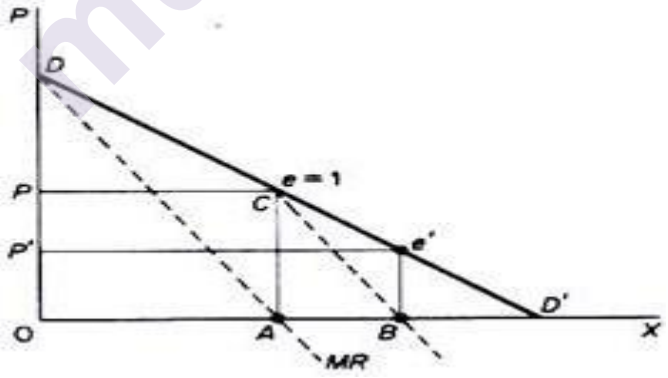
Actually Cournot illustrated his model with the example of two firms each owning a spring of mineral water, which is produced at zero costs. We will present briefly this version, and then we will generalize its presentation by using the reaction curves approach.

Cournot assumed that there are two firms each owning a mineral well, and operating with zero costs. They sell their output in a market with a straight-line demand curve. Each firm acts on the assumption that its competitor will not change its output, and decides its own output so as to maximize profit.

Assume that firm A is the first to start producing and selling mineral water. It will produce quantity A, at price P where profits are at a maximum, because at this point MC — MR = 0. The elasticity of market demand at this level of output is equal to unity and the total revenue of the firm is a maximum. With zero costs, maximum R implies maximum profits, Π. Now firm B assumes that A will keep its output fixed (at 0/1), and hence considers that its own demand curve is CD'.

Clearly firm B will produce half the quantity AD', because (under the Cournot assumption of fixed output of the rival) at this level (AB) of output (and at price F) its revenue and profit is at a maximum. B produces half of the market which has not been supplied by A, that is, B's output is ¼ (= ½. ½) of the total market.

**Fig No. 3.1**




Firm A, faced with this situation, assumes that B will retain his quantity constant in the next period. So he will produce one-half of the market which is not supplied by B. Since B covers one-quarter of the market, A will, in the next period, produce ½(1 – ¼) = ½. ¾ = ⅜ of the total market.

Firm B reacts on the Cournot assumption, and will produce one-half of the unsupplied section of the market, i.e. ½ (1 – ⅜) = 5/16.

In the third period firm A will continue to assume that B will not change its quantity, and thus will produce one-half of the remainder of the market, i.e. ½ (1 – 5/16).

This action-reaction pattern continues, since firms have the naive behaviour of never learning from past patterns of reaction of their rival. However, eventually equilibrium will be reached in which each firm produces one-third of the total market. Together they cover two-thirds of the total market. Each firm maximises its profit in each period, but the industry profits are not maximised.

That is, the firms would have higher joint profits if they recognised their interdependence, after their failure in forecasting the correct reaction of their rival. Recognition of their interdependence (or open collusion) would lead them to act as 'a monopolist,' producing one-half of the total market output, selling it at the profit-maximising price P, and sharing the market equally, that is, each producing one-quarter of the total market (instead of one-third).

**1.3.2 The equilibrium of the Cournot firms may be obtained as follows:**

**1. The product of firm A in successive period is**

**Period 1 :** $\frac{1}{2}$

**Period 2 :** $\frac{1}{2}\left(1-\frac{1}{4}\right)=\frac{3}{8}=\frac{1}{2}-\frac{1}{8}$

**Period 3 :** $\frac{1}{2}\left(1-\frac{5}{16}\right)=\frac{11}{32}=\frac{1}{2}-\frac{1}{8}-\frac{1}{32}$

**Period 4 :** $\frac{1}{2}\left(1-\frac{42}{128}\right)=\frac{43}{128}=\frac{1}{2}-\frac{1}{8}-\frac{1}{32}-\frac{1}{128}$

We observe that the output of A declines gradually. We may rewrite this expression as follows

$$\left[\begin{matrix}Product\ of\ A\\ in\ euilibrum\end{matrix}\right]=\frac{1}{2}-\frac{1}{8}-\frac{1}{32}-\frac{1}{128}\cdots\cdots$$

$$=\frac{1}{2}-\left[\frac{1}{8}+\frac{1}{8}-\frac{1}{4}+\frac{1}{8}.\left(\frac{1}{4}\right)^{2}+\frac{1}{8}.\left(\frac{1}{4}\right)^{3}+\cdots\right]$$

The expression in parentheses is a declining geometric progression with ratio 4=1/4. Applying the summation formula for an infinite geometric series

$$\int\frac{a}{1-r}$$

(Where ∫= sum, a= first term of series, r=ratio) we obtain

$$\left[\begin{matrix}Product\ of\ A\\ in\ euilibrum\end{matrix}\right]=\frac{1}{2}-\frac{\frac{1}{8}}{1-\frac{1}{4}}=\frac{1}{2}-\frac{\frac{1}{8}}{\frac{3}{2}}=\frac{1}{2}-\frac{4}{24}=\frac{8}{24}=\frac{1}{3}$$

2. The product of firm B in successive period is

**Period 2 :** $\frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{4}$

**Period 3 :** $\frac{1}{2}\left(1 - \frac{1}{8}\right) = \frac{5}{16} = \frac{1}{4} + \frac{1}{16}$

**Period 3 :** $\frac{1}{2}\left(1 - \frac{11}{32}\right) = \frac{21}{64} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64}$

**Period 4 :** $\frac{1}{2}\left(1 - \frac{43}{128}\right) = \frac{85}{256} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256}$

We observe that the output of B increases. But at a declining rate We may write

$$\begin{bmatrix} Product\ of\ B \\ in\ euilibrum \end{bmatrix} = \frac{1}{4} + \frac{1}{4}.\frac{1}{4} + \frac{1}{4}.\left(\frac{1}{4}\right)^2 + \frac{1}{4}.\left(\frac{1}{4}\right)^3 + \cdots$$

**Applying the above expression for the summation of a declining geometric series we find**

$$\begin{bmatrix} Product\ of\ B \\ in\ euilibrum \end{bmatrix} = \frac{\frac{1}{4}}{1 - \frac{1}{4}} = \frac{\frac{1}{4}}{\frac{3}{2}} = \frac{1}{3}$$

Thus the Cournot solution is stable. Each firm supplies 4 of the market, at a common price which is lower than the monopoly price, but above the pure competitive price (which is zero in the Cournot example of costless production). It can be shown that if there are three firms in the industry, each will produce one-quarter of the market and all of them together will supply ¾ ( = ¼ . 3) of the entire market OD'.

And, in general, if there are n firms in the industry each will provide n /(n + 1) of the market, and the industry output will be n/(n + 1) = 1 /(n + 1) . n. Clearly as more firms are assumed to exist in the industry, the higher the total quantity supplied and hence the lower the price. The larger the number of firms the closer is output and price to the competitive level.
Cournot's model leads to a stable equilibrium. However, his model may be criticized on several accounts

The behavioural pattern of firms is naive. Firms do not learn from past miscalculations of competitors' reactions.

Although the quantity produced by the competitors is at each stage assumed constant, a quantity competition emerges which drives P down, towards the competitive level.

## 3.3 BERTRAND'S DUOPOLY MODEL

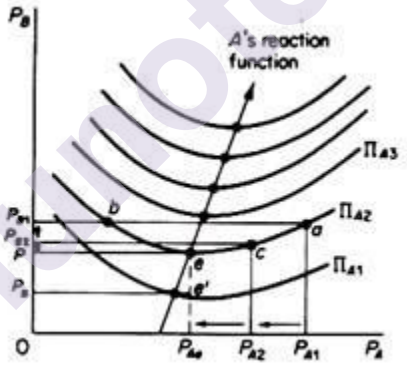
Bertrand developed his duopoly model in 1883. His model differs from Cournot's in that he assumes that each firm expects that the rival will keep its price constant, irrespective of its own decision about pricing.

Thus each firm is faced by the same market demand, and aims at the maximization of its own profit on the assumption that the price of the competitor will remain constant. The model may be presented with the analytical tools of the reaction functions of the duopolist.

In Bertrand's model the reaction curves are derived from iso profit maps which are convex to the axes, on which we now measure the prices of the duopolist. Each iso profit curve for firm A shows the same level of profit which would accrue to A from various levels of prices charged by this firm and its rival.

The iso profit curve for A is convex to its price axis ($P_A$). This shape shows the fact that firm A must lower its price up to a certain level to meet the cutting of price of its competitor, in order to maintain the level of its profits at $\Pi_{A2}$. However, after that price level has been reached and if B continues to cut its price, firm A will be unable to retain its profits, even if it keeps its own price unchanged (at $P_{Ae}$). If, for example, firm B cuts its price at $P_B$, firm A will find itself at a lower iso profit curve ($\Pi_{A1}$) which shows lower profits. The reduction of profits of A is due to the fall in price, and the increase in output beyond the optimal level of utilization of the plant with the consequent increase in costs. Clearly the lower the iso profit curve, the lower the level of profits.

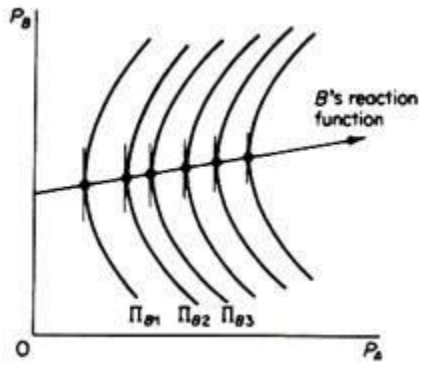
**Figure No. 3.2**




To summaries for any price charged by firm B there will be a unique price of firm A which maximizes the latter's profit. This unique profit-maximizing price is determined at the lowest point on the highest attainable iso profit curve of A. The minimum points of the iso profit curves lie to the right of each other, reflecting the fact that as firm A moves to a higher level of profit, it gains some of the customers of B when the latter increases its price, even if A also raises its price.

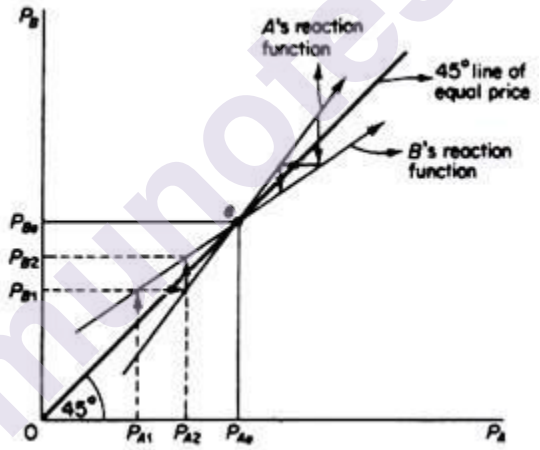
If we join the lowest points of the successive iso profit curves we obtain the reaction curve (or conjectural variation) of firm A: this is the locus of points of maximum profits that A can attain by charging a certain price, given the price of its rival. The reaction curve of firm B may be derived in a similar way, by joining the lowest points of its iso profit curves.

**Figure No. 3.3**




Bertrand's model leads to a stable equilibrium, defined by the point of intersection of the two reaction curves. Point e denotes a stable equilibrium, since any departure from it sets in motion forces which will lead back to point e at which the price charged by A and B are $P_{Ae}$ and $P_{Be}$ respectively. For example, if firm A charges a lower price $P_{A1}$, firm B will charge $P_{B1}$, because on the Bertrand assumption, this price will maximize B's profit (given $P_{A1}$).

**Figure No. 3.4**




Firm A will react to this decision of its rival by charging a higher price $P_{A2}$. Firm B will react by increasing its price, and so on, until point e is reached, when the market will be in equilibrium. The same equilibrium will be reached if firms started by charging a price higher than $P_{Ae}$ or $P_{Be}$ a competitive price cut would take place which would drive both prices down to their equilibrium level $P_{Ae}$ and $P_{Be}$.

Note that Bertrand's model does not lead to the maximization of the industry (joint) profit, due to the fact that firms behave naively, by always assuming that their rival will keep its price fixed, and they never learn from past experience which showed that the rival did not in fact keep its price constant.

**3.3.1 Bertrand's model may be criticised on the same grounds as Cournot's model:**

The behavioural pattern emerging from Bertrand's assumption is naive: firms never learn from past experience. Each firm maximises its own profit, but the industry (joint) profits are not maximized.

The equilibrium price will be the competitive price. (In the example of costless mineral-water production, the price in Bertrand's model would fall to zero. If production is not costless, then price would fall to the level which would cover the costs of the duopolist inclusive of a normal profit.) The model is 'closed'-does not allow entry. The interesting feature of both Cournot's and Bertrand's models is that the limit of duopoly is pure competition. Neither model refutes the other. Each is consistent and is based on different behavioural assumptions. We may say that Bertrand's assumption (about the fixity of price of the rival) is more realistic, in view of the observed preoccupation of firms with keeping their prices constant (except in cost inflation situations).

Furthermore, Bertrand's model focused attention on price setting as the main decision of the firm. The serious limitations of both models are the naive behavioural pattern of rivals; the failure to deal with entry; the failure to incorporate other variables in the model, such as advertising and other selling activities, location of the plant, and changes in the product. Product differentiation and selling activities are the two main weapons of non-price competition, which is a main form of competition in the real business world; both models do not define the length of the adjustment process. Although dealing in terms of 'time periods,' their approach is basically static; both models assume that the market demand is known with accuracy; both models are based on individual demand curves which are located by making the convenient assumption of constant reaction curves of the competing firms.

Having discussed the classical duopoly models of Cournot and Bertrand, we proceed with the development of the traditional models of non-collusive oligopoly, which apply to market structures with a few firms conscious of their interdependence. It is worthwhile pointing out, however, that both Cournot's and Bertrand's models can be extended to markets in which the number of firms is greater than two.

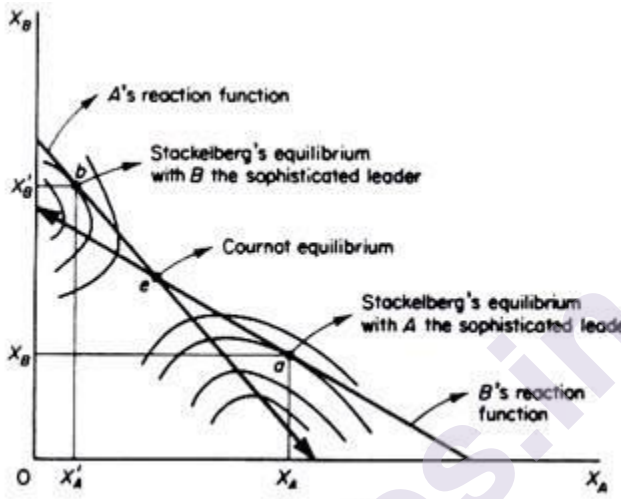
## 3.4 STACKELBERG'S DUOPOLY MODEL

This model was developed by the German economist Heinrich von Stackelberg's and is an extension of Cournot's model. It is assumed, by von Stackelberg's, that one duopolist is sufficiently sophisticated to recognise that his competitor acts on the Cournot assumption.

This recognition allows the sophisticated duopolist to determine the reaction curve of his rival and incorporate it in his own profit function, which he then proceeds to maximise like a monopolist.

Assume that the iso profit curves and the reaction functions of the duopolists are those depicted in following figure. If firm A is the sophisticated oligopolist, it will assume that its rival will act on the basis of its own reaction curve. This recognition will permit firm A to choose to set its own output at the level which maximizes its own profit. This is point a which lies on the lowest possible iso profit curve of A, denoting the maximum profit A can achieve given B's reaction curve.

**Figure No. 3.5**




Firm A, acting as a monopolist (by incorporating B's reaction curve in his profit-maximizing computations) will produce $X_A$, and firm B will react by producing $X_B$ according to its reaction curve. The sophisticated oligopolist becomes in effect the leader, while the naive rival who acts on the Cournot assumption becomes the follower.

Clearly sophistication is rewarding for A because he reaches an iso profit curve closer to his axis than if he behaved with the same naiveté as his rival. The naive follower is worse off as compared with the Cournot equilibrium, since with this level of output he reaches an iso profit curve further away from his axis.

If firm B is the sophisticated oligopolist, it will choose to produce $X'_B$, corresponding to point b on X's reaction curve, because this is the largest profit that B can achieve given his iso profit map and A's reaction curve. Firm B will now be the leader while firm A becomes the follower. B has a higher profit and the naive firm A has a lower profit as compared with the Cournot equilibrium.

In summary, if only one firm is sophisticated, it will emerge as the leader, and a stable equilibrium will emerge, since the naive firm will act as a follower.

However, if both firms are sophisticated, then both will want to act as leaders, because this action yields a greater profit to them. In this case the market situation becomes unstable. The situation is known as

Stackelberg's disequilibrium and the effect will either be a price war until one of the firms surrenders and agrees to act as follower, or a collusion is reached, with both firms abandoning their naive reaction functions and moving to a point closer to (or on) the Edge-worth contract curve with both of them attaining higher profits. If the final equilibrium lines on the Edge-worth contract curve the industry profits (joint profits) are maximised.

It shows clearly that naive behaviour does not pay. The rivals should recognise their interdependence. By recognizing the other's reactions each duopolist can reach a higher level of profit for himself. If both firms start recognising their mutual interdependence, each starts worrying about the rival's profits and the rival's reactions. If each ignores the other, a price war will be inevitable, as a result of which both will be worse off.

The model shows that a bargaining procedure and a collusive agreement become advantageous to both duopolist. With such a collusive agreement the duopolist may reach a point on the Edge-worth contract curve, thus attaining joint profit maximisation.

It should be noted that Stackelberg's model of sophisticated behaviour is not applicable in a market in which the firms behave on Bertrand's assumption. In a Cournot-type market the sophisticated firm 'bluffs' the rival, by producing a level of output larger than the one that would be produced in the Cournot equilibrium and the naive rival, sticking to his Cournot behavioural reaction pattern, will be misled and produce less than in the Cournot equilibrium.

However, in a Bertrand-type market the sophisticated duopolist can do nothing which would increase his own profit and persuade the other to stop price-cutting. The most he can do is to keep his own price constant, that is, behave exactly as his opponent expects him to behave.

A numerical example of Stackelberg's model
Assume that in a Duopoly market the demand function is

$$P = 100 - 0.5(X_1 + X_2)$$

And the Duopolists' costs are

$$C_1 = 5X_1 \; and \; C_2 = 0.5X_2^2$$

The reaction functions are found by taking the partial derivatives of the duopolists profit functions and equating them to zero:

A numerical example of Stackelberg's model

Assume that in a duopoly market the demand function is

$$P = 100 - 0.5(X_1 + X_2)$$

And the duopolists costs are

$$C_1 = 5X_1 \quad and \quad C_2 = 0.5X_2^1$$

The reaction functions are found by taking the partial derivatives of the duopolists profit functions and equating them to zero:

$$\pi_1 = PX_1 - C_1 = 95X_1 - 0.5X_1^2 - 0.5X_1X_2$$

$$\pi_2 = PX_2 - C_2 = 100X_2 - X_1^2 - 0.5X_1X_2$$

The partial derivatives are

$$\frac{\partial \pi_1}{\partial X_1} = 95 - X_1 - 0.5X_2 = 0$$

$$\frac{\partial \pi_2}{\partial X_2} = 100 - 2X_2 - 0.5X_1 = 0$$

The reaction functions are

$$X_1 = 95 - 0.5X_2 \rightarrow A's\ reaction\ curve$$

$$X_2 = 50 - 0.25X_1 \rightarrow B's\ reaction\ curve$$

1] Stackelberg's solution with A being the sophisticated leader

Firm A will substitute B's reaction function in its own profit equation which it will then maximize as if were a monopolist:

$$\pi_1 = PX_1 - C_1 = 95X_1 - 0.5X_1^2 - 0.5X_1X_2$$

Substitute $X_2 = 50 - 0.25X_1$

Maximise $\pi_1 = 70X_1 - 0.375X_1^2$

First order condition: $\frac{\partial \pi_1}{\partial X_1} = 70 - 0.75X_1 = 0$

This yields Output: $X_1 = 93\frac{1}{3}$

And profit : $\pi_1 = 70X_1 - 0.375X_1 = 3267$

(b) The second order condition for profit maximization is fulfilled.

Firm B would be the follower. It would assume that A would produce $93\frac{1}{3}$ units; thus B substitutes this amount in its reaction function

$$X_1 = 50 - 0.25X_1 = 26\frac{2}{3}$$

And its profit would be

$$\pi_2 = 100X_2 - X_2^2 - 0.5X_1X_2 = 155.5$$

(2) Stackelberg's solution if firm B is the sophisticated duopolist

Firm B will substitute A's reaction function in its own profit function, and it will proceed to maximize this profit as a monopolist

$$\pi_2 = PX_2 - C_2 = 100X_2 - X_2^2 - 0.5X_1X_2$$

Substitute $X_1 = 95 - 0.5X_2$ (i.e., A's reaction function)

$$\pi_2 = 52.5X_2 - 0.75X_2^2$$

a) The first order condition for the maximization of $\pi_2$ requires

$$\frac{\partial \pi_2}{\partial X_2} = 52.5 - 1.5X_2 = 0$$

Which yields Output: $X_2 = 35$

And Profit: $\pi_2 = 52.5X_2 - 0.75X_2^2 = 918.75$

(b) The second order condition for the maximization of $\pi_2$ is fulfilled.

The follower is now firm A which will act on the Cournot assumption; it will assume that the rival will keep his quantity at $X_2 = 35$, and will find its own output by substituting this quantity in its reaction function.

$$X_1 = 95 - 0.5X_2 = 77.5$$

And its profit is

$$\pi_1 = 95X_1 - 0.5X_1^2 - 0.5X_1X_2 = 3003$$

**3) Stackelberg's disquilibrium**

If both entrepreneurs adopt Stackelberg's sophisticated pattern of behaviour, each will examine his profits if he acts as a leader and if he acts as a follower, and will adopt the action that will yield him the greatest profit.

**Firm A calculates its profits both as a leader and as a follower:**
If A is the leader his profits are 3267
If A is the follower his profits are 3003
Clearly firm A will prefer to act as the leader.

**Firm B similarly, calculates its profits as a leader and as a follower:**
If B is the leader his profits are 918-75
If B acts as the follower his profits are 155-50
Thus firm B will also choose to act as the leader.

With both firms acting in the sophisticated way implied by Stackelberg's behavioural hypothesis both will want to act as leaders. As they attempt to do so they find that their expectations about the rival are not fulfilled and 'warfare' will start, unless they decide to come to a collusive agreement.

We may now summarise Stackelberg's model. Each duopolist estimates the maximum profit that he would earn (a) if he acted as leader, (b) if he acted as follower, and chooses the behaviour which yields the largest maximum.

**Four situations may arise:**

(1) Duopolist A wants to be leader and B wants to be follower.

(2) Duopolist B wants to be leader and A wants to be follower.

(3) Both firms want to be followers.

(4) Both firms desire to be leaders.

In situations (1) and (2) the result is a determinate equilibrium (provided that the first- and second-order conditions for maxima are fulfilled).
If both firms desire to be followers, their expectations do not materialize (since each assumes that the rival will act as a leader), and they must revise them. Two behavioural patterns are possible. If each duopolist recognises that his rival wants also to be a follower, the Cournot equilibrium is reached. Otherwise, one of the rivals must alter his behaviour and act as a leader before equilibrium is attained.

Finally, if both duopolist want to be leaders disequilibrium arises, whose outcome, according to Stackelberg's, is economic warfare? Equilibrium will be reached either by collusion, or after the 'weaker' firm is eliminated or succumbs to the leadership of the other.

## 3.5 QUESTIONS

Q1. Explain the meaning and features of oligopoly.

a) Types of Oligopoly Market.

Q2. Describe the Cournot's Duopoly Model

Q3. Bertrand's Duopoly Model

Q4. Stackelberg's Duopoly Model

## 3.6 REFERENCES


- H.L AHUJA, Modern Microeconomics, Delhi, S Chand Company Ltd.
- A. Koutsoyiannis (1979), Modern Microeconomics, London – 0-333-77821-9, Macmillan Press Ltd.
- Walter Nicholas & Christopher Snyder, (2008) Micro Economic Theory – Basic Principles and Extension, USA – 10:0-324-42162-1, Thomson South – Western.


*****

# 4

# OLIGOPOLY MODELS - II

**Unit Structure**



## 4.0 OBJECTIVES

- To understand Game theories and its Policies.
- To understand profit maximisation under imperfect competitions.
- To understand the differences in Monopoly and Oligopoly

## 4.1 OLIGOPOLY: REPEATED GAMES

In *game theory*, repeated games, also known as super games, are those that play out over and over for a period of time, and therefore are usually represented using the *extensive form*. As opposed to one-shot games, repeated games introduce a new series of incentives: the possibility of cooperating means that we may decide to compromise in order to carry on receiving a payoff over time, knowing that if we do not uphold our end of the deal, our opponent may decide not to either. Our offer of cooperation or our threat to cease cooperation has to be credible in order for our opponent to uphold their end of the bargain. Working out whether credibility is merited simply involves working out what weighs more: the payoff we stand to gain if we break our pact at any given moment and gain an exceptional, one off payoff, or continued cooperation with lower payoffs which may or may not add up to more over a given time. Therefore, each player must consider their opponent's possible punishment strategies.

This means that the strategy space is greater than in any regular *simultaneous* or *sequential game*. Each player will determine their strategies or moves taking into account all previous moves up until that moment. Also, since each player will take into account this information,

they will play the game based on the behaviour of the opponent, and therefore must consider also possible changes in the behaviour of the latter when making choices.

Repeated games provide different payoffs at each repetition, depending on each player's moves. Since these payoffs are given at different points in time, in order to analyse repeated games, we must compare each player's discounted sum of payoffs, which for infinite repetitions and finite repetitions are calculated using the following formulae:

$$P = \sum_{t=0}^{\infty} \frac{p_t}{(1+r)^t} \qquad P = \sum_{t=0}^{n} \frac{p_t}{(1+r)^t}$$

Where: -P: the discounted sum of payoffs;
-t: the number of the repetition being considered;
-n: the total number of repetitions for finite repeated games;
-$p_t$: the payoff at the repetition being considered;
-r: the discount rate.

**4.1.1 Repeated Prisoner's dilemma:**

In the game known as the *Prisoner's dilemma*, the *Nash equilibrium* is Confess-Confess (defect-defect). In order to see what equilibrium will be reached in a repeated game of the prisoner's dilemma, we must analyse two cases: the game is repeated a finite number of times and the game is repeated an infinite number of times.

When the prisoners know the number of repetitions, it's interesting to operate a backwards induction to solve the game. Consider the strategies of each player when they realise the next round is going to be the last. They behave as if it was a one-shot game, thus the Nash equilibrium applies, and the equilibrium would be confess-confess, just like in the one-time game. Now consider the game before the last. Since each player knows in the next, final round they are going to confess, there's no advantage to lie (cooperate with each other) on this round either. The same logic applies for prior moves. Therefore, confess-confess is the Nash equilibrium for all rounds.

The situation with an infinite number of repetitions is different, since there will be no last round, backwards induction reasoning does not work here. At each round, both prisoners reckon there will be another round and therefore there are always benefits arising form the cooperate (lie) strategy. However, prisoners must take into account punishment strategies, in case the other player confesses in any round.

**4.1.2 Collusion agreement games:**

If we assume the game can be played ad infinitum, we can apply it in a *collusion agreement* game, where two firms collude, forming a cartel. Consider two firms (a *duopoly*) that may either behave as *Cournot*

*duopolist* earning profits $\pi_{Cournot}$ each, or collude and act as a cartel, earning $\pi_{Cartel}$ each, which correspond to the profits of a *monopoly* divided into the number of firms colluding (two in our example).

In this case, we simply need to apply the formula for calculating an infinite sequence and a discount factor to compensate for the fact that the gains to be derived are over time (accounting for impatience, *inflation*, loss of interest, etc.):

$$\Pi$$

$$\frac{\pi^{monopoly}}{2} * \frac{1}{1-\delta} \geq \pi^{deviation} + \frac{\delta}{1-\delta} * \pi^{Cournot}$$

The left hand side represents the payoff derived from collusion, which can be held infinitely over time, with δ being the discount factor to bring future benefits forward to the present. For our threats or offers to be credible, this left hand side must be greater than the right hand side, which represents the one off payoff to be gained from deviating and breaking our cartel. The higher δ is, the higher the value assigned to future benefits and therefore the greater the chances of collusion. It is worth reminding here that fair competition is regulated in almost all countries, with cartels being banned, so most markets that lend themselves to reduced competition and price fixing are closely monitored.

Although this example is widely used in game theory and for the analysis of *market structures*, it can be easily seen that it does not represent a real situation. Let's consider the same example: any of the colluding firms might deviate, in order to dump more output in the market at lower prices, in order to gain market share. This move will allow that firm to sell more products than the other firms, which directly contradicts *Cournot*'s premise that each duopolist will produce the same quantity. Therefore, considering a *Stackelberg duopoly* might seem more realistic. This would obviously change the analysis and outcome of the game.



## 4.2 COMPARISON WITH MONOPOLY

A monopoly and an oligopoly are market structures that exist when there is imperfect competition. A monopoly is when a single company produces goods with no close substitute, while an oligopoly is when a small number of relatively large companies produce similar, but slightly different goods. In both cases, significant barriers to entry prevent other enterprises from competing.

A market's geographical size can determine which structure exists. One company might control an industry in a particular area with no other alternatives, though a few similar companies operate elsewhere in the country. In this case, a company may be a monopoly in one region, but operate in an oligopoly market in a larger geographical area.

**Monopoly:**
A monopoly exists in areas where one company is the only or dominant force to sell a product or service in an industry. This gives the company enough power to keep competitors away from the marketplace. This could be due to high barriers to entry such as technology, steep capital requirements, government regulation, patents or high distribution costs.

Once a monopoly is established, lack of competition can lead the seller to charge high prices. Monopolies are price makers. This means they determine the cost at which their products are sold. These prices can be changed at any time. A monopoly also reduces available choices for buyers. The monopoly becomes a pure monopoly when there is absolutely no other substitute available.

Monopolies are allowed to exist when they benefit the consumer. In some cases, governments may step in and create the monopoly to provide specific services such as a railway, public transport or postal services. For example, the United States Postal Service enjoys a monopoly on first class mail and advertising mail, along with monopoly access to mailboxes.

## 4.3 OLIGOPOLY

In an oligopoly, a group of companies (usually two or more) controls the market. However, no single company can keep the others from wielding significant influence over the industry, and they each may sell products that are slightly different.

Prices in this market are moderate because of the presence of competition. When one company sets a price, others will respond in fashion to remain competitive. For example, if one company cuts prices, other players typically follow suit. Prices are usually higher in an oligopoly than they would be in perfect competition.

Because there is no dominant force in the industry, companies may be tempted to collude with one another rather than compete, which keeps non-established players from entering the market. This cooperation makes them operate as though they were a single company.

In 2012, the U.S. Department of Justice alleged that Apple (AAPL) and five book publishers had engaged in collusion and price fixing for e-books. The department alleged that Apple and the publishers conspired to raise the price for e-book downloads from $9.99 to $14.99. A U.S. District Court sided with the government, a decision which was upheld on appeal.
In a free market, price fixing—even without judicial intervention—is unsustainable. If one company undermines its competition, others are forced to quickly follow. Companies that lower prices to the point where they are not profitable are unable to remain in business for long. Because

of this, members of oligopolies tend to compete in terms of image and quality rather than price.

## 4.4 LEGALITIES OF MONOPOLIES VS. OLIGOPOLIES

Oligopolies and monopolies can operate unencumbered in the United States unless they violate anti-trust laws. These laws cover unreasonable restraint of trade; plainly harmful acts such as price fixing, dividing markets and bid rigging; and mergers and acquisitions (M&A) that substantially lessen competition.

Without competition, companies have the power to fix prices and create product scarcity, which can lead to inferior products and services and higher costs for buyers. Anti-trust laws are in place to ensure a level playing field.

In 2017, the U.S. Department of Justice filed a civil antitrust suit to block AT&T's merger with Time Warner, arguing the acquisition would substantially lessen competition and lead to higher prices for television programming. However, a U.S. District Court judge disagreed with the government's argument and approved the merger, a decision that was upheld on appeal.

The government has several tools to fight monopolistic behaviour. This includes the Sherman Antitrust Act, which prohibits unreasonable restraint of trade, and the Clayton Antitrust Act, which prohibits mergers that lessen competition and requires large companies that plan to merge to seek approval in advance. Anti-trust laws do not sanction companies that achieve monopoly status via offering a better product or service, or though uncontrollable developments such as a key competitor leaving the market.

## 4.5 EXAMPLES OF MONOPOLIES AND OLIGOPOLIES

A company with a new or innovative product or service enjoys a monopoly until competitors emerge. Sometimes these new products are protected by law. For example, pharmaceutical companies in the U.S. are granted 20 years of exclusivity on new drugs. This is necessary due to the time and capital required to develop and bring new drugs to market. Without this protected status, firms would not be able to realize a return on their investment, and potentially beneficial research would be stifled.

Gas and electric utilities are also granted monopolies. However, these utilities are heavily regulated by state public utility commissions. Rates are often controlled, along with any rate increases the company may pass onto consumers.

Oligopolies exist throughout the business world. A handful of companies control the market for mass media and entertainment. Some of the big

names include The Walt Disney Company (DIS), Viacom CBS (VIAC) and Comcast (CMCSA). In the music business, Universal Music Group and Warner Music Group have a tight grip on the market.

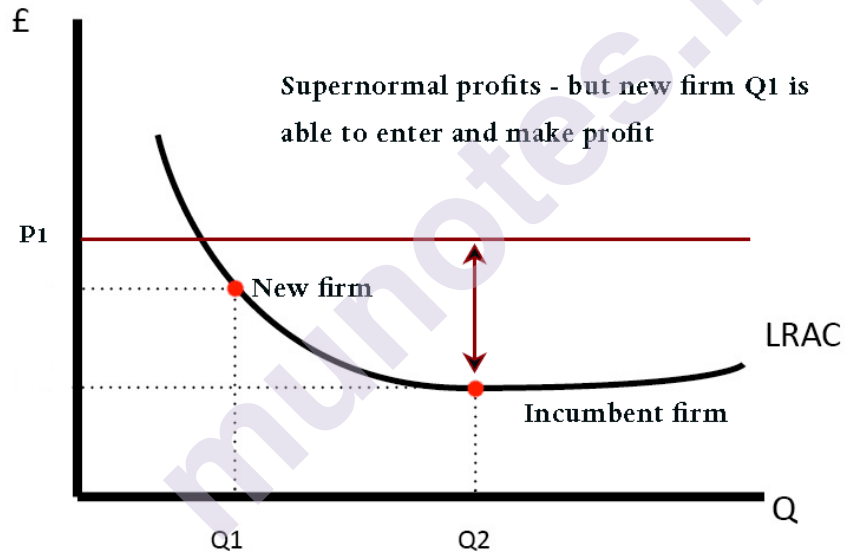
## 4.6 LIMIT PRICING AND ENTRY DETERRENCE IN MONOPOLY

Limit Pricing is a pricing strategy a monopolist may use to discourage entry. If a monopolist set its profit maximising price (where MR=MC) the level of supernormal profit would be so high it attracts new firms into the market. Limit pricing involves reducing the price sufficiently to deter entry. It leads to less profit than possible in short-term, but it can enable the firm to retain its monopoly position and long-term profitability.

### 4.6.1 Profit maximisation in the short run:

In the short-run, a firm may set price using usual profit maximisation rules (where MR=MC). This could lead to a price of P1.

**Figure No. 4.1**



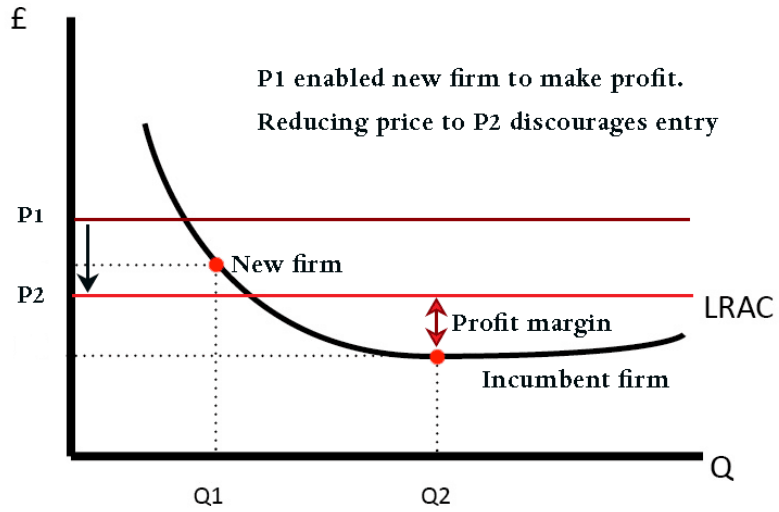

If the new firm produces at Q1, with a market price of P1, that is higher than its average costs – so it is profitable for a new firm to enter.

### 4.6.2 Limit pricing:

Therefore, rather than encouraging a new firm to enter, the monopolist may decide to set a price below this profit maximising level, but still high enough to enable it to make higher profits than in a competitive market.

For limit pricing to be effective, the monopolist needs to decrease the price to the point where a new firm will not be able to make any profit on entering the market.

**Figure No. 4.2**




By reducing the price to P2 it sacrifices supernormal profit in the short-run. But, the price is low enough to discourage a new firm entering. At P2, a new firm faces average costs higher than the market price.

By discouraging entry, the incumbent firm is guaranteed an 'easy life' and guaranteed high profits. The monopolist may also build excess capacity as a threat that if firms enter, it will reduce the price even further.

### 4.6.3 Evaluation of limit pricing:

- A large multinational may be willing to enter a market – even if it is unprofitable in the short-term. The large multinational can use its reserves and profit elsewhere to subsidies a loss-making entry. For example, Google entered the market for mobile phones – despite no experience. Limit pricing is not effective if new firms have the capacity to absorb losses.
- Rather than limit pricing, a firm may set the profit maximising price, but then react when a new firm enters. If a new firm enters, it lowers price to make it difficult. It could go to an extreme and engage in predatory price – setting the price below average cost to force the rival out of business. Predatory pricing is illegal, which is a reason to choose limit pricing instead.
- Limit pricing will be more effective in industries with substantial economies of scale – for example, industries, such as steel and aeroplane manufacture. It gives an advantage to the incumbent and disadvantage to potential new firms. For industries, with few economies of scale, such as restaurants and bars, limit pricing will not be effective

## 4.7 QUESTIONS

**Explain**

1. Oligopoly: Repeated games.
2. Repeated Prisoner's dilemma.
3. Comparison with monopoly.
4. Limit Pricing and Entry Deterrence in Monopoly

## 4.8 REFERENCES


- H.L AHUJA , Modern Microeconomics, Delhi, S Chand Company Ltd.
- A. Koutsoyiannis (1979), Modern Microeconomics, London – 0-333-77821-9, Macmillan Press Ltd.
- Robert Pendyck & Daniel Rubinfeld, 2017, Microeconomics , 13-978-9332585096, Pearson Publication.
- Andreu, Michael & Jerry, 2012, Microeconomic Theory, Oxford Publication.


*****

# MODULE III

# 5

# MORAL HAZARD AND ADVERSE SELECTION-I

**Unit Structure**



## 5.0 OBJECTIVES

- To understand the concepts of moral hazard and adverse selection.
- To know the market for lemons
- To understand the principal agent models
- To understand the efficiency wage/effort model.

## 5.1 MORAL HAZARD

If a person is not insured for life or property, he would be more careful with his life and property and avoid taking risks with his life and property. However, when he ensures his life and property, his behavior becomes more risky, thereby exposing him to premature death or loss of property. **The tendency of insured persons to be more prone to risk and thereby increasing the probability of the insured event happening is known as moral hazard.** A person who has insured his house against fire and theft would be less careful about his property. Similarly, a person who has insured his car against theft would not think twice before parking his vehicle in a public place. He may also have no incentive in obtaining a car park in his residential premises. Further a person whose car is stolen would not make all the necessary efforts to obtain his car back for he is sure that the value of the car would be paid to him by the insurance company. These are all examples of moral hazard and it is because of the problem of moral hazard that insurance companies do not offer insurance premiums at fair odds. The insurance companies would try to reduce the problem of moral hazard by offering conditional coverage. For instance,

the insurance company may cover a residential house or a firm's premises only if fire detection and fire fighting system is installed. In case of health insurance, the insurance company insists on medical check-up for identifying any pre-existing diseases and any such disease is not covered by the policy. In this way, the insurance companies are able to charge small premiums and reduce claims.

The insurance companies need to find out the optimal combination of premium and risk covered. Let us assume that a person who insures his house against fire. The value of his house is W and if fire occurs the value of his house will be reduced to $W_2$ ($W_2 = W - D$, here D = debris). The individual insures his house against fire by paying a premium $\infty_1$. The house is insured against fire for amount equal to $\infty_2$. If there is no fire, his wealth is $W_1 = W - \infty_1$ and if there is fire his wealth is $W_2 = W - d + \infty_2$.

**Figure No. 5.1**
**The Problem of Moral Hazard.**

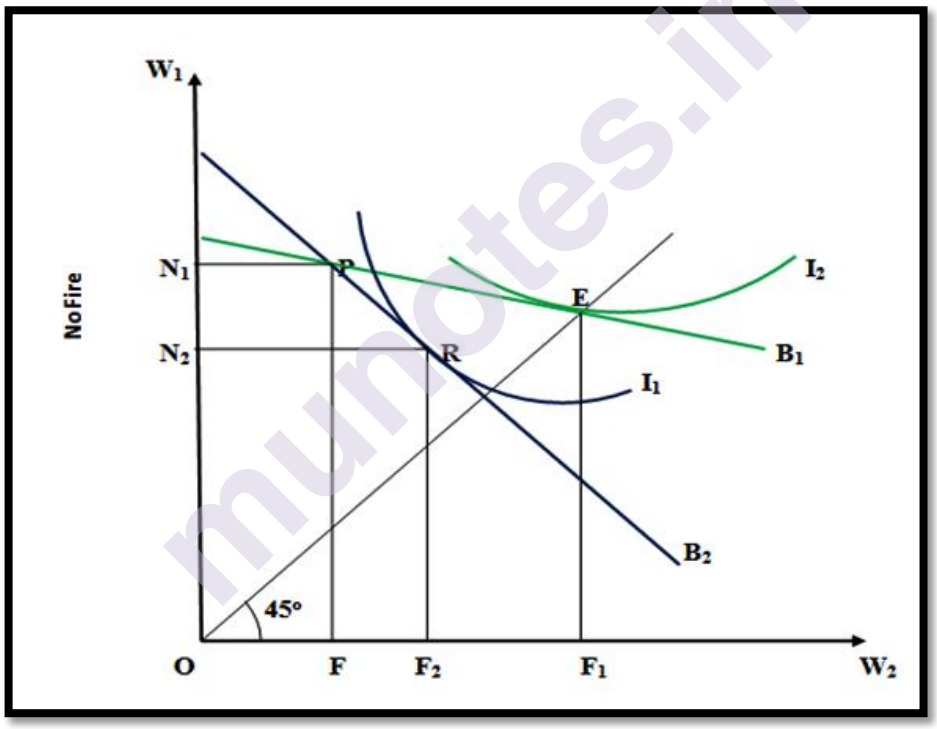



**Fig.5.1 - The Problem of Moral Hazard.**

Insurance companies expose themselves to lesser risks and hence offer less favorable odds to their customers. By offering less favorable odds, insurance companies are also able to reduce the problem of moral hazard. This is shown in Figure 5.1 Let us begin with point P which represents the value of the house of the person. In the absence of insurance, the value of his house will be reduced to OF in the event of fire. Let us also assume that the probability of 'no fire' is three times the probability of the house going on fire i.e. 3 to 1. This is shown by the slope of the person's budget line $B_1$ whose slope is 1/3 indicating 3 to 1 odds. Let us now assume that

the householder insures his house against fire. Assuming that a fire occurs with probability of 1 to 3, he chooses point E where his budget line $B_1$ and indifference curve $I_1$ are tangent. Point E is the risk free point for the householder which is along the $45^o$ line because by paying $\infty_1 = NN_1$ insurance premium his wealth remains $W_1 = W - \infty_1$ or $ON_1 = OF_1$ whether there is fire or no fire. Therefore, he will not take precautions against fire and hence a fire is most likely to occur. You may note that along the $45^o$ line, $W_2 = W$ or $W - d + \infty_2 = W_1 - \infty_1$. Hence the payment by the insurance company just covers the loss of house in case of a fire. The insurance company will therefore not offer 3 to 1 odds. Being a risk-averse organization, it will sell the insurance policy at much less than the full value of the house to safeguard itself against loss due to moral hazard and also laying down certain conditions in the policy. This situation is shown in Fig.3.8 where the householder's equilibrium point is R where his budget line is $B_2$ and the indifference curve $I_2$ are tangent to each other. At point R, the householder is paying the same premium $NN_1$ but in case of fire, he will be paid the insured sum $OF_2$ instead of $OF_1$ of the earlier insured amount.

## 5.2 ADVERSE SELECTION

Adverse selection takes place when customers know more than the insurance company about the probability of an event happening. For instance, in the market for individual health insurance, the person who seeks a health insurance cover knows more about his health issues than the insurance company. In order to cover the risk of inadequate information, the insurance company will charge a premium based on the national average. This will dissuade healthy persons from taking up health insurance cover because they think that the premium is unreasonably high and more unhealthy persons will buy insurance cover because they think that the premium is low. As a result, high-risk individuals are more likely to buy insurance than low-risk individuals. **This problem is known as the problem of adverse selection.** Adverse selection has the potential to bankrupt an insurance company and hence insurance companies may hike the premium to a level such that even unhealthy persons may not buy insurance cover. Insurance companies solve the problem of adverse selection by charging different premiums for different age groups and occupation based on the nature of risk in each group. Thus low risk groups would be charged low premiums and high risk groups would be charged high premiums. Persons in different age groups are charged different rates of premium depending on the length of the period of insurance and the risk involved.

## 5.3 THE MARKET FOR LEMONS

Real life is imperfect and full of uncertainties. Uncertainties involve risks. There are political, social, economic and natural uncertainties. Uncertainties are always unforeseen and there is no way that uncertainties

can be prevented from happening. These uncertainties are not factored in the basic economic theories. All units in an economy have to confront uncertainties. The households, the firms and the government constitute the three basic units of the macro-economy. Households worry about future income flows, wages and employment. They may worry about the return on their investments in financial markets. The employment market and the capital market trends particularly the stock market trends cannot be accurately predicted. The labor skills demanded in future may be entirely different from the kind of skills imparted and acquired by students in our colleges and universities. The stock market may be rising in a sustained manner but suddenly may fall like a pack of cards and the wealth owned by a large number of people may evaporate overnight. **The study of unforeseen factors is known as the economics of uncertainties and risk.**

**Asymmetric information** explains situations in which not all individuals involved in a potential exchange are equally well informed. Generally, the seller of a product or a service has more knowledge about the quality of the product or the service than the potential buyers. Asymmetric information prevents mutually beneficial exchange in the markets for high quality goods because buyers do not have adequate information to select high quality goods and hence they are not willing to pay a fair price. Asymmetric information and other communication problems between potential exchange partners can be generally solved through the use of signals that are costly or difficult to fake. For instance, product warranties are such a signal. The seller of a low quality product would not offer a product warranty because it would prove costly to him. Buyers and sellers may react to asymmetric information by attempting to judge the qualities of products and people on the basis of the groups to which they belong. For example, a young taxi driver knows that he is a good driver but the insurance company would still charge him a high premium because the taxi driver is a member of a group that is frequently involved in accidents.

**The Problem of Asymmetric Information:**

Sandeep has a well maintained Maruti Baleno Car and now he decided to buy a trendy car. He wants to sell his car. The current market price for 2013 Maruti Suzuki Alto car is Rs. Two lakh sixty five thousand but Sandeep wants to sell his car for Rs.Three lakhs because he knows that his car is in good condition. Sanjay wants to buy an old Maruti Suzuki Alto Car and would be willing to pay Rs.Three lakh fifty thousand for a car in good condition but only Rs. Two lakh seventy five thousand for a car in not so good condition. Sandeep can hire a mechanic to assess the condition of the car. However, not all the faults can be detected by a mechanic. Will Sanjay buy Sandeep's car? Because Sandeep's 2013 Maruti Alto looks no different from other similar cars, Sanjay is not willing to pay Rs.Three lakhs. Sanjay can buy another 2013 Maruti Alto only for Rs.Two lakh fifty thousand which according to him is as good as Sandeep's car. In this situation, Sanjay will buy someone else's car and Sandeep's car will remain unsold. The outcome of this potential exchange

is not efficient. If Sanjay had bought Sandeep's Maruti Alto for Rs.Three lakhs his surplus would have been Rs. Fifty thousand and Sanjay would have got a fair price for his car. But Sanjay goes ahead and buys a Maruti Alto for Rs.Three lakh twenty five thousand. Sandeep's car remains unsold and Sanjay gets only Rs.25 thousand as surplus.

**The Lemons Model:**

It is difficult to conclude that the car that Sanjay ends up buying will be in worse condition than Sandeep's car because anybody else may have a better car than Sandeep and yet may not find a buyer to pay a fair price. However, the economic incentives created by asymmetric information suggest that most used cars that are put up for sale will be of low quality. This is because people who ill-treat their cars or had purchased a not so good car are more likely to sell them than others. Buyers also know from their experience that cars for sale on the used car market are more likely to be **'lemons'** than cars that are not for sale. This realization on the part of buyers causes them to bargain a used car at a lower reservation price. Further when the prices of used cars fall in the market, the owners of cars that are in good condition will not offer their cars for sale. This causes the average quality of the cars offered for sale on the used car market to decline further. **George Akerlof, a Nobel laureate economist from Berkeley**, was the first to explain the logic behind such a price fall. Economists use the term **'lemons model'** to describe Akerlof's explanation of how asymmetric information affects the average quality of the used goods offered for sale.

The lemons model has important practical implications for consumer choice. These implications are exemplified in the following illustrations.

**Should you buy your friend's car?:**

You want to buy a used Hyundai Accent (GLE). Your friend buys a new car every three years and he has a three year old Hyundai Accent (GLE) which he wants to sell. Your friend says that his car is in good condition and he is willing to sell it to you for Rs.3.5 lakhs which is the average market price for three year old Accents. Should you buy your friend's car? Going by the Lemons Model, it would make little sense to buy a used car because used cars offered for sale in the market are of lower quality than those cars of the same vintage not offered for sale. If your friend's claim regarding the condition of the car was to be believed then buying a car at an average price will certainly be a good deal for you because the average price is always a price for a lower quality car than what is claimed by the owner. Illustrations 3.3 and 3.4 will help you understand the conditions under which asymmetric information about the quality of product results in a market in which only poor quality products or lemons are offered for sale.

**What price will an innocent buyer pay for a used car?**

Let us consider a market with only two kinds of cars: lemons and good ones. The owner of a car certainly knows the quality of his car but

potential buyers can in no way distinguish between the lemons and good ones. Ninety per cent of the new cars are good but ten per cent of them turn out to be lemons. Used but good cars are worth Rs. Five lakhs but lemons are worth only Rs. Three lakhs. Let us consider an innocent buyer who thinks that the used cars for sale have the same quality distribution as new cars. Assuming that this buyer is risk-neutral, what price will he be willing to pay for a used car? It is a gamble to buy a car of unknown quality. However, a risk-neutral buyer will be willing to play the gamble if it is a fair gamble. The buyer here is not able to distinguish between lemons and good cars. Yet, given the distribution of good cars and lemons, the buyer has a 90 per cent chance of buying a good car and a ten per cent chance of buying a lemon. Given the prices that he is willing to pay for the two types of car, his expected value of the car he buys will thus be 0.90(Rs.5 Lakhs) + 0.10(Rs.3 Lakhs) = Rs.4.8 lakhs. The buyer is a risk-neutral person and hence his reservation price for a used car will be Rs.4.8 Lakhs.

**Who will sell a used car for a price that an innocent buyer is willing to pay?**

If you are the owner of a used good car, at what price would you be willing to sell your car? Would you sell it to an innocent buyer? What if your car turns out to be a lemon? You know that your car is good and hence it is worth Rs.Five lakhs to you but an innocent buyer will be willing to pay only Rs.4.8 lakhs. Hence, neither you nor anybody else who owns a good car will be willing to sell it for that price. If you had a lemon, you will be all the more happy to sell it to an innocent buyer because Rs.4.8 lakhs that the buyer is willing to pay is Rs.1.8 lakhs more than the lemon's worth to you. Hence only used cars for sale will be lemons. In due course of time, buyers will revise their optimistic beliefs about the quality of the used cars for sale. Finally, all used cars will sell for a price of Rs. Three lakhs and all will be lemons. However, in practice, it does not mean that all cars offered for sale are lemons because the owner of a good car may sell it at an average price under compelling circumstances. The lemons model explains the frustration of such owners. When you buy a used car that is sold for reason that has nothing to do with the condition of the car for an average price, you are actually beating the market i.e. you are buying a good car for the price of a lemon.

**The Problem of Credibility in Trading:**

It is difficult for a seller to convince the buyer about the good quality of the car that he has offered for sale. This difficulty is due to the conflicting interests of the buyers and the sellers. Sellers of used cars have an economic incentive to overstate the quality of their products and buyers have an incentive to understate the amount they are willing to pay for used cars and other products. There is a tendency amongst people to interpret ambiguous information in such a manner that it promotes their self interest. However, both buyers and sellers can gain if they can find some means to communicate their knowledge truthfully. This is described in the following illustration.

**Credible manner in which the good quality of the car can be signaled (The Costly-to-fake Principle):**

Sandeep knows that his car is in good condition and Sanjay would be willing to pay much more than his reservation price for a good car. What kind of signal about the car's quality would Sanjay find credible? The potential conflict of interest between Sandeep and Sanjay shows that mere statements about the quality of the car may not persuade Sanjay to buy the car. Let us suppose that Sandeep offers a warranty under which he agrees to repair any defects the car develops over the next one year. Sandeep can afford to offer such a warranty because he knows his car is unlikely to need expensive repairs. In contrast, the person who knows his car would need extensive repairs would never extend such an offer. The warranty is a credible signal that the car is in good condition. It enables Sanjay to buy the car with confidence and both Sandeep and Sanjay would gain in such a deal.

Illustration 3.5 exemplifies the costly- to-fake principle. This principle suggests that if parties whose interests potentially conflict are to communicate credibly with each other, the signals they send must be costly to fake. Warranties cannot be faked because bad quality products would impose heavy costs on the seller to offer warranties.

## 5.4 THE PRINCIPAL AGENT PROBLEM

The principal agent problem is a situation where the principal due to want of knowledge cannot ensure his best interest is served by the agent. For example, in a class room setting, the students are the principal and the teacher is the agent. Due to want of information, the students are not in a position to know if the teacher is doing his best to serve their interests. In a corporate setting, the principal is the owner and the agent constitutes the managers. The managers may pursue their own goals rather than pursuing the goals of the owners. The principal agent problem is due to the problem of asymmetric information. An agency relationship comes into existence when there is an arrangement in which one person's welfare depends upon what other person does. The agent is the person who acts and the principal is the party whom the action affects. A principal – agent problem arises when agents pursue their own goals and not the goals of the principal.

In a modern economy, principals have to employ agents to carry out their tasks. Whether it is firms or companies and their employees, sick persons and medical doctors, students and teachers, principals and agents have to come together to satisfy their goals. However, due to asymmetric information, it is difficult for the principle to judge in whose interest the agent is operating. The medical doctor may prescribe unnecessary medical examinations or tests, the teacher may not cover the portion entirely and source his information from the prescribed reference books and employees in a firm may shirk from performing expected tasks.

**Measures to Reduce the Principal Agent Problem.**

**1. Performance Monitoring:** The principals must monitor the performance of their agents. In corporate settings, the performance of the employee is monitored and evaluated by the human resource department. Annual increments, promotions and demotions are awarded on the basis of performance evaluation of the employees.

**2. Incentives for Agents:** The principals in any setting must create a system of incentives and disincentives. While incentives will motivate the agents to perform according to the expectations of the principles, disincentives will dissuade the agents from shirking or working below their natural potential.

## 5.5 EFFICIENCY WAGE/EFFORT MODEL

According to **efficiency wage** hypothesis, in some markets, wages are determined by factors other than the market forces of supply and demand. Managers pay their employees more than the market-clearing wage in order to increase their productivity or efficiency which in turn compensates for the higher wages. Since workers are paid more than the market clearing or equilibrium wage, there will be unemployment. Efficiency wages are therefore a market failure explanation of unemployment which is in contrast to theories which emphasize government intervention such as minimum wages. The idea of efficiency wages was expressed as early as 1920 by Alfred Marshall. Efficiency wage theory is especially important in new Keynesian economics. Theories which explain as to why managers pay efficiency wages are:

**1. Avoiding Shirking**. If it is difficult to measure the quantity or quality of a worker's effort and systems of piece rates or commissions are impossible. There may be an incentive for the employee to 'shirk' i.e. to do less work than agreed. The manager thus may pay an efficiency wage in order to create or increase the opportunity cost, which gives the threat of firing. This threat can be used to prevent shirking (or moral hazard).

**2. Minimizing Turnover:** The worker's motivation to leave the job and look for a job elsewhere will be reduced due to efficiency wages. Efficiency wages makes economic sense because it is often expensive to train replacement workers.

**3. Adverse Selection:** Firms with higher wages will attract more able job-seekers. An efficiency wage means that the employer can choose the best workers among applicants, thus eliminating the problem of adverse selection.

**4. Sociological Theories:** Efficiency wages may result from traditions. According to Akerlof's theory, higher wages leads to high morale and high productivity.

**5. Nutritional Theories**. In developing countries, efficiency wages may allow workers to eat well enough to avoid illness and to be able to work harder and more productively.

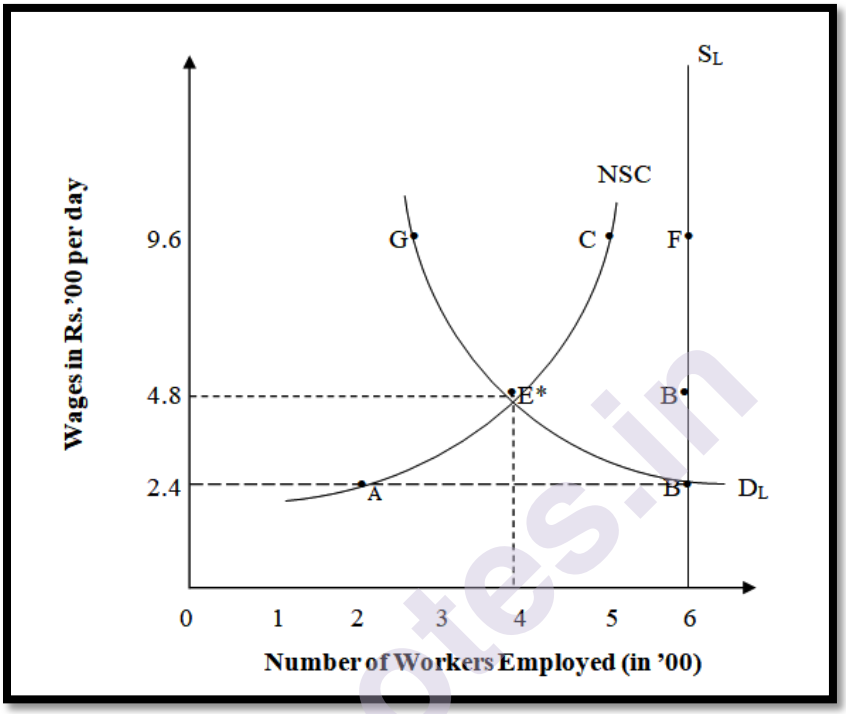
**The Theory:**

According to the Efficiency Wage theory, firms willing to pay higher than equilibrium wages to workers do so to incentivize them to avoid shirking. It is impossible to monitor workers' productivity correctly. Firms therefore face a principal-agent problem caused by asymmetric information. Since there is no involuntary unemployment at the equilibrium wage rate, workers who are fired for shirking will easily find re-employment. Hence, by paying an efficiency wage which is higher than the equilibrium wage, the firm can induce workers to work without shirking and with greater effort and productivity. There is always an opportunity cost in losing a high paying job. Further, at efficiency wage rates, there is substantial involuntary unemployment and if workers shirk at efficiency wage rate, they will be fired with no chances of being reemployed at higher than equilibrium wage rates. The theory can be explained with figure 5.2 below.

In Figure 5.2, $D_L$ is the demand curve for labor by the firm and $S_L$ is assumed to be perfectly inelastic labor supply curve at the equilibrium wage rate determined at point 'E' which is the intersection point between the downward sloping demand curve and the vertically sloping supply curve of labor. Here, OW is the equilibrium wage rate and OL is the equilibrium demand and supply of labor. At Rs.240/- per day wage rate, there is no involuntary unemployment and is equal to marginal productivity of labor (Rs.240 = $MP_L$). However, at Rs.240 wage rate, workers have a tendency to shirk. In order to prevent shirking, firms will have to pay a wage rate which is higher than the equilibrium wage rate. The higher the efficiency wage, the smaller is the level of unemployment. This is shown by the no-shirking constraint (NSC) curve. The NSC curve shows the minimum wage that workers must be paid for each level of unemployment to avoid shirking. For instance, at the efficiency wage of Rs.240, the number of unemployed workers would be EA. When the efficiency wage is raised to Rs.480 per day, the number of workers unemployed is only BE* and when CF workers are unemployed the efficiency wage is Rs. 960/- per day. The NSC curve is positively sloped i.e. smaller the level of unemployment, higher will be the efficiency wage. The NSC curve will neither intercept nor intersect the $S_L$ curve because there will be some unemployment at the efficiency wage. The intersection point between $D_L$ and NSC is point E* where the efficiency wage determined is Rs.480 per day. At this wage rate, the firm employs 400 workers and 200 workers remains unemployed. Unemployment of 200 workers is considered enough prevent shirking amongst the employed workers at the wage rate of Rs.480 per day. At a lower efficiency wage rate of Rs.240 per day, the number of workers required to be unemployed is 300 (EA). However, at this wage rate, the unemployment is zero (point E). Hence, the equilibrium efficiency wage rate must be higher.

Conversely, at Rs.960 per day wage rate, only 100 workers need to be unemployed (FC) but the actual unemployment is of 350 workers (FG). Hence, the equilibrium wage rate must be lower. The efficiency wage is Rs.580 because at the wage rate, the number of unemployed workers (200) is just good enough to prevent the employed workers from shirking.

**Fig.No. 5.2 - Efficiency Wage and Unemployment (Shirking Model).**




**The Carl Shapiro & Joseph Stiglitz Model Of Efficiency Wages:**

In the Shapiro-Stiglitz model workers are paid at a level where they do not shirk. This prevents wages from dropping to equilibrium or market clearing levels. Full employment cannot be achieved because workers would shirk if they were not threatened with the possibility of unemployment. According to the shirking model, complete contracts do not exist in the real world. This implies that both parties to the contract have some discretion, but frequently, due to monitoring problems, it is the employee's side of the bargain which is subject to the most discretion. Methods such as piece rates are impracticable because monitoring is too costly or inaccurate. Such methods may be based on measures too imperfectly verifiable by workers, creating a moral hazard problem on the employer's side. Thus the payment of a wage in excess of market-clearing may provide employees with cost-effective incentives to work rather than shirk.

In the Shapiro and Stiglitz model, workers either work or shirk and if they shirk they have a certain probability of being caught with the penalty of being fired. As a result, at the point of equilibrium there is unemployment. Unemployment is generated because firms try to push their wages above the market average to create an opportunity cost to shirking. This creates a low, or no income alternative which makes job loss costly, and serves as

an instrument of discipline for the workers. Unemployed workers cannot bid for jobs by offering to work at lower wages, since if hired, it would be in the worker's interest to shirk on the job, and he or she has no credible way of promising not to shirk. Shapiro and Stiglitz point out that their assumption that workers are identical (e.g. there is no stigma to having been fired) is a strong one – in practice reputation can work as an additional disciplining device.

The shirking model does not predict that the bulk of the unemployed at any one time are those who are fired for shirking, because if the threat associated with being fired is effective, little or no shirking and sacking will occur. Instead the unemployed will consist of a rotating pool of individuals who have quit for personal reasons, are new entrants to the labor market, or who have been laid off for other reasons. Pareto optimality, with costly monitoring, will result in some unemployment, since unemployment plays a socially valuable role in creating work incentives. But the equilibrium unemployment rate will not be Pareto optimal, since firms do not take into account the social cost of the unemployment they help to create.

However, the efficiency wage hypothesis is criticized on the ground that more sophisticated employment contracts can under certain conditions reduce or eliminate involuntary unemployment. Lazear demonstrates the use of seniority wages to solve the incentive problem, where initially workers are paid less than their marginal productivity, and as they work effectively over time within the firm, earnings increase until they exceed marginal productivity. The upward bias in the age-earnings profile provides the incentive to avoid shirking, and the present value of wages can fall to the market-clearing level, eliminating involuntary unemployment. Lazear and Moore find that the slope of earnings profiles is significantly affected by incentives.

However, a significant criticism is that moral hazard would be shifted to employers, since they are responsible for monitoring the worker's effort. Incentives would exist for firms to declare shirking when it has not taken place. In the Lazear model, firms have incentives to fire older workers (paid above marginal product) and hire new cheaper workers, creating a credibility problem. The seriousness of this employer moral hazard depends on the extent to which effort can be monitored by outside auditors, so that firms cannot cheat, although reputation effects may have the same impact.

## 5.6 QUESTIONS

Q1. Write an explanatory note on moral hazard and adverse selection.

Q2. Write a note on the market for lemons.

Q3. What is the Principal Agent problem? How the problem is solved by the Efficiency Wage model?

*****

# 6

# MORAL HAZARD AND ADVERSE SELECTION-II

**Unit Structure**



## 6.0 OBJECTIVES

- To help the learner with the clear understanding of the concept of Screening and signalling.
- How these two concepts are used at different cases such as the selection procedure of the candidate or the labourer, his promotion etc.
- Student will also learn that how these concepts are used by the insurance companies.

## 6.1 INTRODUCTION

Asymmetric information exists when one party in a transaction possesses better information than the other party. In certain industries, some parties in a transaction are bound to know more than other parties in the same transaction.

For example, in a sale transaction, sellers are bound to have more information than the buyers, since dealing with the same product or a range of products gives them greater knowledge of the product compared to the knowledge that some buyers have.

Screening and Signalling is used when asymmetric information may lead to a moral hazard or adverse selection due to information imbalance.

## 6.2 SCREENING

Screening in economics refers to a strategy of combating adverse selection – one of the potential decision-making complications in cases of asymmetric information – by the agent(s) with less information.

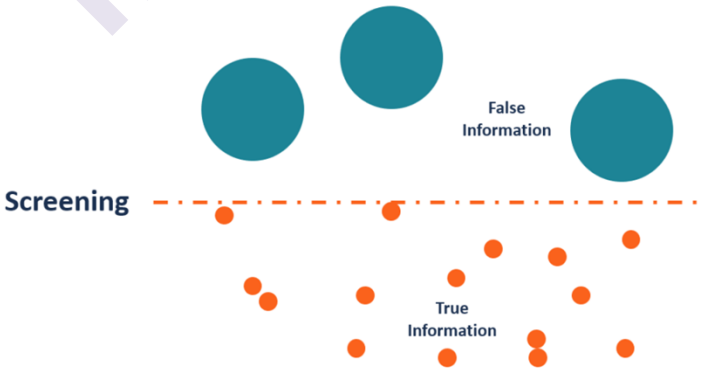
Screening refers to a strategy that is used to combat adverse selection by filtering out false information and retaining only the true information. Screening is used in contemporary markets where the products being released into the market are getting increasingly complex for an ordinary consumer to comprehend.

For the purposes of screening, asymmetric information cases assume two economic agents, with agents attempting to engage in some sort of transaction. There often exists a long-term relationship between the two agents, though that qualifier is not necessary. Fundamentally, the strategy involved with screening comprises the "screener" (the agent with less information) attempting to gain further insight or knowledge into private information that the other economic agent possesses which is initially unknown to the screener before the transaction takes place. In gathering such information, the information asymmetry between the two agents is reduced, meaning that the screening agent can then make more informed decisions when partaking in the transaction.

Screening is applied in a number of industries and markets. The exact type of information intended to be revealed by the screener ranges widely; the actual screening process implemented depends on the nature of the transaction taking place. Often it is closely connected with the future relationship between the two agents.

The concept of screening was first developed by Michael Spence (1973)

**Fig. No. 6.1**




For example, in the auto industry, non-specialist buyers rely on the information provided by the seller when evaluating the type of car they want to buy. Since the specialist seller possesses more information than the buyer, he or she may give false information about a product in order to convince the buyer to purchase that item instead of another. Screening is

employed in various areas, such as insurance, job markets, and management, where the problem of asymmetric information exists.

### 6.2.1 Screening Techniques in Labour Market:

Screening theory provides an alternative with regard to education, production and wages. As hypothesized by Spence (1973), Arrow (1973), and Stiglitz (1975), it proclaimed education to be an essential screen or signal to productivity. According to Brown and Sessions, higher education is viewed as an endorsement to perform higher-level jobs yielding higher wages. Proponents of the screening theory maintain that it provides the optional explanation that links organizational behaviors with the labor market .

Screening theory addresses the selection needs of organizations in order to make ideal hiring decisions that yield desired production requirements. Thus, the theory considers the function that education plays in communicating necessary information to organizations and assumes that employers first establish the required education levels that classify job applicants. Education acts as a screening mechanism that signals an individual's capabilities. Completion of education and training programs are often requirements or prerequisites to promotions and other personnel decisions. Degrees and diplomas indicate employee production potential. Organizations can obtain education information in a low-cost manner to use in hiring decisions.

Employees with higher levels of education have certain characteristics that include favorable attendance records and less likelihood of engaging in unhealthy habits such as smoking, excessive drinking, and illicit drug use. screening theory acknowledges the positive correlation between education and wages. The screening theory argues that employers operate in imperfect labor markets and employees utilize the various general and specific skills during the process of performing the duties and expectations required by organizations.

Screening techniques are employed within the labour market during the hiring and recruitment stage of a job application process. In brief, the hiring party (agent with less information) attempts to reveal more about the characteristics of potential job candidates (agents with more information) so as to make the most optimal choice in recruiting a worker for the role.

There are several techniques that employers use to address the problem of asymmetric information among interview candidates. Screening Techniques Used in the Labor Market are:

1. **Application Review:** The hiring party initially screens applicants by undertaking a review of their application submission and any responses received, including an evaluation of their resume and cover letter to reveal education, experience and fit for the role

2. **Aptitude Tests and Assessment:** Aptitude tests are one of the most popular screening techniques that employers use to select high-quality candidates from a pool of job seekers. Aptitude tests are usually in the form of specialized tests that are used to test a candidate's productivity and their knowledge of specific subjects. The hiring party may require applicants to undertake a range of testing exercises (either online or in-person) to reveal academic or practical abilities
3. **Quality of College or University:** Employers also use the candidate's school affiliation to shortlist candidates. They assume that the top-tier colleges and universities produce high-quality candidates who are likely to outperform candidates from the other colleges.
4. **Grade Point Average (GPA):** The average grade points achieved during the years spent in school can also be used to screen potential employees. The top performers who have performed consistently well in school will have high averages compared to students who have had varied performances during their years in school.
5. **Interviews:** Candidates are often required to undertake an interview with a representative(s) from the hiring party to reveal a range of factors such as personality traits, verbal communication ability and confidence level.

There are many examples of screening in employment decisions. Employers give aptitude tests and check letters of recommendation. The existence of "old-boy" networks is the result of a screening process. If a person wants to hire someone, he will ask those he trusts (the "old boys") for recommendations. Because recommending someone who is unqualified will lower his prestige in the eyes of the other "old boys," there is an incentive for a person to only recommend qualified applicants.1 Also, part of the enthusiasm that employers have for graduates of prestigious MBA programs is that the schools are selective about who they let in. They try to select only those students who have the right combinations of intelligence and personality traits to ensure success in the business world. Thus, prestigious MBA programs act as a screening agency for business. This, as much as what they teach their students, may account for the high salaries their graduates command.

### 6.2.2 Screening Techniques in Insurance Market:

The best known theoretical explanation is that of competitive screening, put out by Rothschild and Stiglitz in 1977, in their article "Equilibrium in Competitive Insurance Markets", which shows how insurance companies can get around the people taking advantage of adverse selection by offering different types of insurance options which will attract only the risk adverse. This is covered in more detail in insurance models, which are covered by the field analysing risk and uncertainty.

**There are two basic types of screening:** in the first, the 'victim' of asymmetric information simply sets about finding out as much as possible about the other agent. For example, carrying out a health check before

offering health insurance, or running a background check before offering a job. This, aside from verging on the morally questionable, is often highly regulated in many countries. The second option is using game theory to set up the terms of a contract so that they only interest the cherries. Something as simple as copayment in case of a claim (for example, paying a small percentage of the claim amount in case a car is damaged) can help to weed out those who are not risk adverse.

The process of screening customers is highly applicable in the market for insurance. In general, parties providing insurance perform such activities to reveal the overall risk level of a customer, and as such, the likelihood that they will file for a claim. When in possession of this information, the insuring party can ensure a suitable form of cover (i.e. commensurate with the customer's risk level) is provided. Asymmetric information also exists in the insurance industry, and it often leads to moral hazard among the insured persons. Some of the techniques that insurance companies use include:

**1. Historical record:**
Insurers look at the past behavior of its insurance clients to determine their level of risk and the possibility that they will engage in risky behavior in the future. Background check is undertaken by the party providing insurance to obtain information about the customer such as their criminal history, credit rating and previous employment to reveal past behaviors. For example, if a client has a history of multiple car accidents in the past, there is a likelihood that the client will still get involved in an accident in the future. It makes the insurance company aware of the level of risk that it is subjecting itself to by providing insurance coverage to the risky client.

**2. Health condition:**
When providing life insurance coverage to a client, the insurer will be interested in knowing the health condition of the client and the kind of illnesses that the person has. Clients with terminal illnesses or other long-term illnesses are usually categorized as risky and are, therefore, charged different premiums compared to clients with no history of illnesses.

3. Demographic Characteristics or Provision of Demographic Information
Another consideration that insurance companies make is looking at the demographic characteristics of its new clients. The party providing insurance obtains information about the customer such as their age, gender and ethnicity to reveal their type When selling auto insurance, younger clients in the 13- to 20-year-old bracket are considered risky compared to the clients in the 40- to 50-year-old age bracket. On the other hand, older clients aged above 60 years old are considered risky compared to younger clients aged 30 to 40 years old in life insurance.

Other information gathered by insurance parties during a screening process is usually specific to the type of insurance the customer is seeking. For example, car insurance will require provision of accident history,

health insurance will require provision of health condition and previous illnesses, and so on.

When there is asymmetric information in the market, screening can involve incentives that encourage the better informed to self-select or self-reveal. For example, a job with a low-paying probationary period will discourage those who know they are not well-suited for the position from applying. People who are confident that they will survive the probationary period are more likely to find the offer attractive than those who doubt their ability. A lender who demands collateral for a loan discourages applications from those who doubt their ability to repay. (Collateralized loans do more than screen, but screening is one of their functions.) People who expect to use insurance find deductibles more of a burden than those who do not expect to make claims. Hence, insurance companies use deductibles to sort policyholders into different risk classes and charge accordingly.

**6.2.3 Other Techniques:**

Second degree price discrimination is also an example of screening, whereby a seller offers a menu of options and the buyer's choice reveals their private information. Specifically, such a strategy attempts to reveal more information about a buyer's willingness to pay. For example, an airline offering economy, premium economy, business and first class tickets reveals information regarding the amount the customer is willing to spend on their airfare. With such information, firms can capture a greater portion of total market surplus.

In contract theory, the terms "screening models" and "adverse selection models" are often used interchangeably. An agent has private information about his type (e.g., his costs or his valuation of a good) before the principal makes a contract offer. The principal will then offer a menu of contracts in order to separate the different types. Typically, the best type will trade the same amount as in the first-best benchmark solution (which would be attained under complete information), a property known as "no distortion at the top". All other types typically trade less than in the first-best solution (i.e., there is a "downward distortion" of the trade level).

Optimal auction design (more generally known as Bayesian mechanism design) can be seen as a multi-agent version of the basic screening model. Contract-theoretic screening models have been pioneered by Roger Myerson and Eric Maskin. They have been extended in various directions. For example, it has been shown that, in the context of patent licensing, optimal screening contracts may actually yield too much trade compared to the first-best solution. Applications of screening models include regulation, public procurement, and monopolistic price discrimination. Contract-theoretic screening models have been successfully tested in laboratory experiments and using field data.

## 6.3 MARKET SIGNALLING

Signalling theory is fundamentally concerned with reducing information asymmetry between two parties. In contract theory, signalling is the idea that one party (termed the agent) credibly conveys some information about itself to another party (the principal). Although signalling theory was initially developed by Michael Spence based on observed knowledge gaps between organisations and prospective employees, its intuitive nature led it to be adapted to many other domains, such as Human Resource Management, business, and financial markets.

In Spence's job-market signalling model, (potential) employees send a signal about their ability level to the employer by acquiring education credentials. The informational value of the credential comes from the fact that the employer believes the credential is positively correlated with having the greater ability and difficulty for low ability employees to obtain. Thus, the credential enables the employer to reliably distinguish low ability workers from high ability workers. The concept of signalling is also applicable in competitive altruistic interaction, where the capacity of the receiving party is limited.

Signalling started with the idea of asymmetric information (a deviation from perfect information), which relates to the fact that, in some economic transactions, inequalities exist in the normal market for the exchange of goods and services. In his seminal 1973 article, Michael Spence proposed that two parties could get around the problem of asymmetric information by having one party send a signal that would reveal some piece of relevant information to the other party. That party would then interpret the signal and adjust his or her purchasing behaviour accordingly—usually by offering a higher price than if she had not received the signal. There are, of course, many problems that these parties would immediately run into.

- How much time, energy, or money should the sender (agent) spend on sending the signal?
- How can the receiver (the principal, who is usually the buyer in the transaction) trust the signal to be an honest declaration of information?
- Assuming there is a signalling equilibrium under which the sender signals honestly and the receiver trusts that information, under what circumstances will that equilibrium break down?

Suppose that Jeevan wants to sell a car that he values at Rs.50000/-. Hari is looking for a car and would consider Jeevan's car worth Rs.60000/- if he knew as much about it as Jeevan knows. An exchange would benefit both Hari and Jeevan but it might not take place because of an information problem. Jeevan probably knows a variety of things about his car that might not be obvious to a buyer. But how can Hari trust Jeevan to tell him all that he knows when Jeevan has the incentive to misrepresent the quality of the car?

Economists say that the potential transaction described above has the problem of asymmetric information, which simply means that the information available to buyers is different than the information available to sellers. They are interested in this problem because they see it in many different situations and because it may lead to a market failure, a case in which a market is economically inefficient. However, when there is unexploited value, buyers and sellers have an incentive to find ways to capture that value. Sellers with high quality products need ways to signal the quality of their products so that buyers can distinguish between high-quality and low-quality products. Buyers must find ways to screen out erroneous information but allow in truthful information. These problems do not exist in markets in which products are simple and easily evaluated. There is little need for this behavior in many agricultural markets, for instance.

One way a seller can signal the quality of its product is by offering guarantees or warranties. If a firm offers a warranty on a poor product, it will suffer a loss. Therefore, it is in the firm's interests to only offer a warranty on a quality product. The warranty tells potential buyers that the firm will stake money on its belief that it has a good-quality product.

Another way a firm can signal quality is by building a brand name. A brand name is valuable only if consumers associate it with quality, and the firm can build this association only with time and resources. Once a brand name is established, it is in the interests of the firm to protect it by not offering a poor-quality product with its brand name. When a firm with an established brand name does offer a poor-quality product, it usually puts a different name on the product so as not to endanger the public's perception of its brand name.

Signalling plays an important role in the labor market. An employer has little information about a prospective employee, and cannot expect truthful answers if he asks whether the applicant is intelligent, has leadership qualities, and is responsible. Instead, the applicant must try to prove that he has these qualities. A college education is a way of signalling intelligence and perseverance. Leadership can be signalled by extracurricular activities. (As a result, some students seek leadership positions primarily for their value as ways to signal leadership to future employers.) The purpose of a resume is to list those activities that will signal attractive qualities to potential employers.

The fact that a college education can signal qualities to employers has raised some interesting questions about why people get college educations. A popular answer among economists has been that education builds human capital, that is, it is a way of investing in people to increase their productivity. More recently some economists have suggested that this view is wrong or at best only partly true, and that college education mostly serves as a way of signalling to future employers. If education is merely a way of signalling, if it is only a complex gauntlet that eliminates those

who are not intelligent and do not have perseverance, then the social usefulness of college education may not be very great. From the viewpoint of the student, it does not matter--the benefits are the same either way. Although most economists believe that education both builds human capital and acts as a signal, the relative importance of these two functions is still disputed.

Signalling is an action by a party with good information that is confined to situations of asymmetric information. Screening, which is an attempt to filter helpful from useless information, is an action by those with poor information.

Michael Spence considers hiring as a type of investment under uncertainty analogous to buying a lottery ticket and refers to the attributes of an applicant which are observable to the employer as indices. Of these, attributes which the applicant can manipulate are termed signals. Applicant age is thus an index but is not a signal since it does not change at the discretion of the applicant. The employer is supposed to have conditional probability assessments of productive capacity, based on previous experience of the market, for each combination of indices and signals. The employer updates those assessments upon observing each employee's characteristics. The paper is concerned with a risk-neutral employer. The offered wage is the expected marginal product. Signals may be acquired by sustaining signalling costs (monetary and not). If everyone invests in the signal in the exactly the same way, then the signal can't be used as discriminatory, therefore a critical assumption is made: the costs of signalling are negatively correlated with productivity. This situation as described is a feedback loop: the employer updates his beliefs upon new market information and updates the wage schedule, applicants react by signalling, and recruitment takes place. Michael Spence studies the signalling equilibrium that may result from such a situation.

He began his 1973 model with a hypothetical example: suppose that there are two types of employees—good and bad—and that employers are willing to pay a higher wage to the good type than the bad type. Spence assumes that for employers, there's no real way to tell in advance which employees will be of the good or bad type. Bad employees aren't upset about this, because they get a free ride from the hard work of the good employees. But good employees know that they deserve to be paid more for their higher productivity, so they desire to invest in the signal—in this case, some amount of education. But he does make one key assumption: good-type employees pay less for one unit of education than bad-type employees. The cost he refers to is not necessarily the cost of tuition and living expenses, sometimes called out of pocket expenses, as one could make the argument that higher ability persons tend to enroll in "better" (i.e. more expensive) institutions. Rather, the cost Spence is referring to is the opportunity cost. This is a combination of 'costs', monetary and otherwise, including psychological, time, effort and so on. Of key importance to the value of the signal is the differing cost structure between

"good" and "bad" workers. The cost of obtaining identical credentials is strictly lower for the "good" employee than it is for the "bad" employee. The differing cost structure need not preclude "bad" workers from obtaining the credential. All that is necessary for the signal to have value (informational or otherwise) is that the group with the signal is positively correlated with the previously unobservable group of "good" workers. In general, the degree to which a signal is thought to be correlated to unknown or unobservable attributes is directly related to its value.

**The result:**
Spence discovered that even if education did not contribute anything to an employee's productivity, it could still have value to both the employer and employee. If the appropriate cost/benefit structure exists (or is created), "good" employees will buy more education in order to signal their higher productivity.

The increase in wages associated with obtaining a higher credential is sometimes referred to as the "sheepskin effect", since "sheepskin" informally denotes a diploma. It is important to note that this is not the same as the returns from an additional year of education. The "sheepskin" effect is actually the wage increase above what would normally be attributed to the extra year of education. This can be observed empirically in the wage differences between 'drop-outs' vs. 'completers' with an equal number of years of education. It is also important that one does not equate the fact that higher wages are paid to more educated individuals entirely to signalling or the 'sheepskin' effects. In reality, education serves many different purposes for individuals and society as a whole. Only when all of these aspects, as well as all the many factors affecting wages, are controlled for, does the effect of the "sheepskin" approach its true value. Empirical studies of signalling indicate it as a statistically significant determinant of wages, however, it is one of a host of other attributes—age,sex, and geography are examples of other important factors.

**The Model:**
To illustrate his argument, Spence imagines, for simplicity, two productively distinct groups in a population facing one employer. The signal under consideration is education, measured by an index y and is subject to individual choice. Education costs are both monetary and psychic. The data can be summarized as:

**Table No. 6.1**

**Data of the Model**

| Group | Marginal Product | Proportion of population | Cost of education level y |
|---|---|---|---|
| I | 1 | | y |
| II | 2 | | y/2 |

Suppose that the employer believes that there is a level of education y* below which productivity is 1 and above which productivity is 2. His offered wage schedule W(y) will be:

Working with these hypotheses Spence shows that:

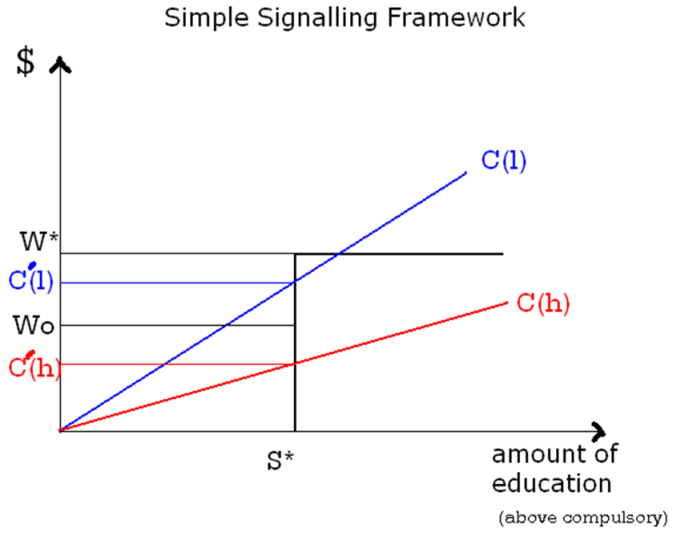
1. There is no rational reason for someone choosing a different level of education from 0 or y*.
2. Group I sets y=0 if 1>2-y*, that is if the return for not investing in education is higher than investing in education.
3. Group II sets y=y* if 2-y*/2>1, that is the return for investing in education is higher than not investing in education.
4. Therefore, putting the previous two inequalities together, if 1<y*<2, then the employer's initial beliefs are confirmed.
5. There are infinite equilibrium values of y* belonging to the interval [1,2], but they are not equivalent from the welfare point of view. The higher y* the worse off is Group II, while Group I is unaffected.
6. If no signalling takes place each person is paid his unconditional expected marginal product. Therefore, Group, I is worse off when signalling is present.

In conclusion, even if education has no real contribution to the marginal product of the worker, the combination of the beliefs of the employer and the presence of signalling transforms the education level y* in a prerequisite for the higher paying job. It may appear to an external observer that education has raised the marginal product of labor, without this necessarily being true.

Another model

For a signal to be effective, certain conditions must be true. In equilibrium, the cost of obtaining the credential must be lower for high productivity workers and act as a signal to the employer such that they will pay a higher wage.

**Fig. No. 6.2**



In this model it is optimal for the higher ability person to obtain the credential (the observable signal) but not for the lower ability individual. The table shows the outcome of low ability person l and high ability person h with and without signal S*:

**Summary of the outcome for l and h with and without $S^*$**

| Person | Without Signal | With Signal | Will the person obtain the signal $S^*$? |
|---|---|---|---|
| l | $W_o$ | $W^* - C'(l)$ | No, because $W_o > W^* - C'(l)$ |
| h | $W_o$ | $W^* - C'(h)$ | Yes, because $W_o < W^* - C'(h)$ |

The structure is as follows: There are two individuals with differing abilities (productivity) levels.

- A higher ability / productivity person: h
- A lower ability / productivity person : l

The premise for the model is that a person of high ability (h) has a lower cost for obtaining a given level of education than does a person of lower ability (l). Cost can be in terms of monetary, such as tuition, or psychological, stress incurred to obtain the credential.

- Wo is the expected wage for an education level less than S*
- W* is the expected wage for an education level equal or greater than S*

**For the individual:**

**$Person_{(credential)} - Person_{(no\ credential)} \geq Cost_{(credential)} \rightarrow$ Obtain credential**

**$Person_{(credential)} - Person_{(no\ credential)} < Cost_{(credential)} \rightarrow$ Do not obtain credential**

Thus, if both individuals act rationally, it is optimal for person h to obtain S* but not for person l so long as the following conditions are satisfied.

note that this is incorrect with the example as graphed. Both `1` and `h` have lower cost than W* at the education elevel. Also, $Person_{(credential)}$ and $Person_{(no\ credential)}$ are not clear.

note that this is ok as for low type "l": and thus, low type will choose Do not obtain credential.

For there to be a separating equilibrium the high type 'h' must also check their outside option; do they want to choose the net pay in the separating equilibrium (calculated above) over the net pay in the pooling equilibrium. Thus, we also need to test that: Otherwise high type 'h' will choose Do not obtain credential of the pooling equilibrium.

For the employers:

$\text{Person}_{(\text{credential})} = E(\text{Productivity} \mid \text{Cost}_{(\text{credential})} \leq \text{Person}_{(\text{credential})} - \text{Person}_{(\text{no credential})})$

$\text{Person}_{(\text{no credential})} = E(\text{Productivity} \mid \text{Cost}_{(\text{credential})} > \text{Person}_{(\text{credential})} - \text{Person}_{(\text{no credential})})$

In equilibrium, in order for the signalling model to hold, the employer must recognize the signal and pay the corresponding wage and this will result in the workers self-sorting into the two groups. One can see that the cost/benefit structure for a signal to be effective must fall within certain bounds or else the system will fail.

**Costly signalling:**

In foreign policy, it is common to see game theory problems such as the prisoner's dilemma and chicken game occur as the different parties both have a dominating strategy regardless of the actions of the other party. In order to signal to the other parties, and furthermore for the signal to be credible, strategies such as tying hands and sinking costs are often implemented. These are examples of costly signals which typically present some form of assurance and commitment in order to show that the signal is credible and the party receiving the signal should act on the information given. Despite this however, there is still much contention as to whether, in practice, costly signalling is effective. In studies by Quek (2016) it was suggested that decision makers such as politicians and leader don't seem to interpret and understand signals the way they that models suggest they should.

**Prisoners Dilemma**

| | B Cooperate | B Defect |
|---|---|---|
| A Cooperate | 3,3 | 0,5 |
| A Defect | 5,0 | 1,1 |

**Chicken's Game**

| | B Swerve | B Don't Serve |
|---|---|---|
| A Swerve | 0,0 | -1,1 |
| A Don't Swerve | 1,-1 | -5,-5 |

**Sinking costs and Tying hands:**

A costly signal in which the cost of an action is incurred upfront ("ex ante") is a sunk cost. An example of this would be the mobilization of an army as this sends a clear signal of intentions and the costs are incurred immediately.

When the cost of the action is incurred after the decision is made ("ex post") it is considered to be tying hands. A common example being an

alliance made which doesn't have a large monetary cost initially however, it does tie the hands of the parties as they are now reliant on one another in a time of crisis.

Theoretically both sinking costs and tying hands are valid forms of costly signalling however they have garnered much criticism due to differing beliefs regarding the overall effectiveness of the methods in altering the likelihood of war. Recent studies such as the Journal of Conflict Resolution suggest that sinking costs and tying hands are both effective in increasing credibility. This was done by finding how the change in the costs of costly signals vary their credibility. Prior to this research studies conducted were binary and static by nature, limiting the capability of the model. This increased the validity of the use of these signalling mechanisms in foreign diplomacy.

**Conclusion:**
Individuals may also serve as their own insiders when signalling about themselves (e.g., in the job market). Signals about investments are also common, but we combine these with organizational signals. Signalling theory focuses mainly on costly signals, scholars have also extended research on information asymmetries to include less costly forms of communication. For example, Farrell and Rabin (1996), in an article titled "Cheap Talk," provided an influential analysis of how insiders communicate cost-less information.

Screening is one of the main strategies for combating adverse selection. It is often confused with signalling, but there is one main difference: in both, 'good' agents (the cherries of this world) are set apart from the 'bad' agents, or lemons, which are weeded out. In signalling, it is the uninformed agent (the victim of asymmetric information) who moves first, and comes up with a strategy to weed out the lemons. In signalling, however, it is the cherries, the informed agents, who make the first move to set themselves apart.

## 6.4 SUMMARY

In this way the concept of screening and Signalling help us to overcome the problems created by lack of information. It helps us to know the specific steps require to be taken in the say labour or insurance market. It is essential for both the parties.

## 6.5 QUESTIONS

Q1. Write a note on the concept of Screening

Q2. Explain the process of screening in labour market.

Q3. How screening and signalling help in an insurance market.

Q4. Explain the concept of Signalling

## 6.6 REFERENCES


- Arrow, K. J. (1971). Essays in the theory of risk bearing. New York: North Holland.
- Gravelle H. and Rees R.(2004) : Microeconomics., 3rd Edition, Pearson Edition Ltd, New Delhi.
- Gibbons R. A Primer in Game Theory, Harvester-Wheatsheaf, 1992
- A. Koutsoyiannis : Modern Microeconomics
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.
- Varian H (2000): Intermediate Microeconomics: A Modern Approach, 8th Edition, W.W.Norton and Company
- Varian: Microeconomic Analysis, Third Edition
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.
- Williamson, O. E. (1988). Corporate finance and corporate governance. Journal of Finance, 43, 567–591.


*****

# MODULE IV

# 7

# ALTERNATIVE THEORIES OF THE FIRM-I

**Unit Structure**



## 7.0 OBJECTIVES

After studying this module, you shall be able to

- Know the concept of managerial theory of firm
- Why Williamson model is different from other managerial theories?
- Williamson's Utility Function
- Understand why utility maximization is the goal of the managers rather than profit maximization?
- Behavioural theories of firm
- Full cost pricing principle
- Existence, purpose and boundaries of firm
- Resource, Knowledge and Transaction cost-based theories of firm

## 7.1 INTRODUCTION

**Alternative Theories of The Firm:**

The traditional theories of firm had analysed the decision-making on the basis of the objective of profit maximisation. As an alternative, Baumol had put forward the notion that the firms maximise sales revenue. Williamson analyses the case for a firm which it maximises the managerial

utility function subject to a profit constraint. Marris in his model shows the equilibrium as a fallout of maximisation of both the owners and the managers

Alternative Theories of the Firm provides a range of fundamental readings embracing the economics of firm behaviour from a non-neoclassical perspective. The collection covers several basic topics including: the importance of transaction costs and agency theory for the analysis of firm behaviour; capabilities and resource-based theories of the firm; the economics of firm strategy; behavioural theories; Austrian theories; evolutionary theories; and the historical development of firms. The readings include selections from traditional masters as well as writings by more recent authors. This collection will be of great value both to scholars who want a summary of developments in the field and to students of industrial economics and corporate strategy.

Managerial theories conceive the firm as a 'coalition' (of managers, workers, stock-holders, suppliers, customers tax collectors) whose members have conflicting goals that must be reconciled if the firm is to survive. The conflicts are resolved by top management by various methods explained in behavioural theories.

It will be studies in two parts. First part will cover the Marris, Williamson models and Behavioural theories of firm. It will also cover the concept of full-cost pricing. In second part we will study Existence, purpose and boundaries of firm and Resource, Knowledge and Transaction cost-based theories of firm

## 7.2 MORRIS/MARRIS MODEL OF MANAGERIAL ENTERPRISE

**I. Goals of the Firm:**

The goal of the firm in Marris's model is the maximisation of the balanced rate of growth of the firm, that is, the maximisation of the rate of growth of demand for the products of the firm, and of the growth of its capital supply:

Maximise g = $g_D$= $g_C$

where g = balanced growth rate

$g_D$= growth of demand for the products of the firm

$g_C$= growth of the supply of capital

In pursuing this maximum balanced growth rate, the firm has two constraints. Firstly, a constraint set by the available managerial team and its skills. Secondly, a financial constraint, set by the desire of managers to achieve maximum job security. These constraints are analysed in a subsequent section. The rationalisation of this goal is that by jointly maximising the rate of growth of demand and capital the managers

achieve maximisation of their own utility as well as of the utility of the owners-shareholders.

It is usually argued by managerial theorists that the division of ownership and manage-ment allows the managers to set goals which do not necessarily coincide with those of owners. The utility function of managers includes variables such as salaries, status, power and job security, while the utility function of owners includes variables such as profits, size of output, size of capital, share of the market and public image. Thus, managers want to maximise their own utility

$U_M$= (salaries, power, status, job security)

while the owners seek the maximisation of their utility

$U_O$= f* (profits, capital, output, market share, public esteem).

Marris argues that the difference between the goals of managers and the goals of the owners is not so wide as other managerial theories claim, because most of the variables appearing in both functions are strongly correlated with a single variable the size of the firm (see below). There are various measures (indicators) of size capital, output, revenue, market share, and there is no consensus about which of these measures is the best.

However, Marris limits his model to situations of steady rate of growth over time during which most of the relevant economic magnitudes change simultaneously, so that 'maximising the long-run growth rate of any indicator can reasonably be assumed equivalent to maximising the long-run rate of most others.' (Marris, 'A Model of the Managerial Enterprise'.) Furthermore, Marris argues that the managers do not maximise the absolute size of the firm (however measured), but the rate of growth (= change of the size) of the firm. The size and the rate of growth are not necessarily equivalent from the point of view of managerial utility. If they were equivalent, we would observe a high mobility of managers between firms: the managers would be indifferent in choosing between being employed and promoted within the same growing firm (enjoying higher salaries, power and prestige), and moving from a smaller firm to a larger firm where they would eventually have the same earnings and status.

In the real world the mobility of managers is low. Various studies provide evidence that managers prefer to be promoted within the same growing organisation rather than move to a larger one, where the environment might be hostile to the 'newcomer' and where he would have to give considerable time and effort to 'learn' the mechanism of the new organisation. Hence managers aim at the maximisation of the rate of growth rather than the absolute size of the firm.

Marris argues that since growth happens to be compatible with the interests of the shareholders in general, the goal of maximisation of the growth rate (however measured) seems a priori plausible. There is no need

to distinguish between the rate of growth of demand (which maximises the u of managers) and the rate of growth of capital supply (which maximises the U of owners) since in equilibrium these growth rates are equal.

From Marris's discussion it follows that the utility function of owners can be written as follows

$U_{owners}$ = f*($g_c$)

where $g_c$= rate of growth of capital.

It is not clear why owners should prefer growth to profits, unless gc and profits are positively related. At the end of his article Marris argues in fact that gc and IT are not always positively related. Under certain circumstances gc and II become competing goals. Furthermore, from Marris's discussion of the nature of the variables of the managerial utility function it seems that he implicitly assumes that salaries, status and power of managers are strongly correlated with the growth of demand for the products of the firm: managers will enjoy higher salaries and will have more prestige the faster the rate of growth of demand. Therefore, the managerial utility function may be written as follows

$U_M$ = f ($g_D$,s)

where $g_D$ = rate of growth of demand for the products of the firm

s = a measure of job security.

Marris, following Penrose, argues that there is a constraint to $g_D$set by the decision-making capacity of the managerial team. Furthermore, Marris suggests that 's' can be measured by a weighted average of three crucial ratios, the liquidity ratio, the leverage- debt ratio and the profit-retention ratio, which reflect the financial policy of the firm.

As a first approximation Marris treats 's' as an exogenously determined constraint by assuming that there is a saturation level for job security above the saturation level the marginal utility from an increase in 's' (job security) is zero, while below the saturation level the marginal utility from an increase in 's' is infinite. With this assumption the managerial utility function becomes

$U_M$ = f ($g_D$,) $\bar{s}$

where $\bar{s}$ is the security constraint. Thus, in the initial model there are two constraints – the managerial team constraint the job security constraint – reflected in a financial constraint. Let us examine these constraints in some detail.

**II. Constraints:**

**The Managerial Constraint:**

Marris adopts Penrose's thesis of the existence of a definite limit on the rate of efficient managerial expansion. At any one time period the capacity of the top management is given there is a ceiling to the growth of the firm

set by the capacity of its managerial team. The managerial capacity can be increased by hiring new managers, but there is a definite limit to the rate at which management can expand and remain competent (efficient).

Penrose's theory is that decision-making and the planning of the operations of the firm are the result of teamwork requiring the co-operation of all managers. Co-ordination and co-operation require experience. A new manager requires time before he is fully ready to join the teamwork necessary for the efficient functioning of the organisation. Thus, although the 'managerial ceiling' is receding gradually, the process cannot be speeded up.

Similarly, the 'research and development' (R & D) department sets a limit to the rate of growth of the firm. This department is the source of new ideas and new products, which affect the growth of demand for the products of the firm. The work in the R & D department is 'teamwork' and as such it cannot be expanded quickly, simply by hiring more personnel for this section: new scientists and designers require time before they can efficiently contribute to the teamwork of the R & D department.

The managerial constraint and the R & D capacity of the firm set limits both to the rate of growth of demand ($g_D$) and the rate of growth of capital supply ($g_c$).

**The Job Security Constraint:**
We said that the managers want job security; they attach (not surprisingly) a definite disutility to the risk of being dismissed. The desire of managers for security is reflected in their preference for service contracts, generous pension schemes, and their dislike for policies which endanger their position by increasing the risk of their dismissal by the owners (that is, the shareholders or the directors they appoint). Marris suggests that job security is attained by adopting a prudent financial policy.

The risk of dismissal of managers arises if their policies lead the firm towards financial failure (bankruptcy) or render the firm attractive to take-over raiders. In the first case the shareholders may decide to replace the old management in the hope that by appointing new management the firm will be run more successfully. In the second case, if the take-over raid is success-ful, the new owners may well decide to replace the old management.

**The risk of dismissal is largely avoided by:**

(a) Non-involvement with risky invest-ments. The managers choose projects which guarantee a steady performance, rather than risky ventures which may be highly profitable, if successful, but will endanger the managers' position if they fail. Thus, the managers become risk-avoiders.

(b) Choosing a 'prudent financial policy'. The latter consists of determining optimal levels for three crucial financial ratios, the leverage (or debt ratio), the liquidity ratio, and the retention ratio.

**The leverage or debt ratio** is defined as the ratio of debt to the gross value of total assets of the firm:

$$\begin{bmatrix}\text{Leverage}\\ \text{or}\\ \text{Debt ratio}\end{bmatrix} = \frac{\text{Value of debts}}{\text{Total assets}} = \frac{D}{A}$$

The managers do not want excessive borrowing because the firm may become insolvent and be proclaimed bankrupt, due to demands for interest payments and repayment of loans, notwithstanding the good prospects that the firm may have.

**The liquidity ratio** is defined as the ratio of liquid assets to the total gross assets of the firm:

$$\begin{bmatrix}\text{Liquidity}\\ \text{ratio}\end{bmatrix} = \frac{\text{Liquid assets}}{\text{Total assets}} = \frac{L}{A}$$

Liquidity policy is very important. Too low a liquidity ratio increases the risk of in-solvency and bankruptcy. On the other hand, too high a liquidity ratio makes the firm attractive to take-over raids, because the raiders think that they can utilise the excessive liquid assets to promote the operations of their enterprises. Thus the managers have to choose an optimal liquidity ratio neither too high nor dangerously low. In his model, however, Marris assumes without much justification, that the firm operates in the region where there is a positive relation between liquidity and security: an increase in liquidity increases security.

**The retention ratio** is defined as the ratio of retained profits (net of interest on debt) to total profits:

$$[\text{Retention ratio}] = \frac{\text{Retained profits}}{\text{Total profits}} = \frac{\Pi_R}{\Pi}$$

Retained profits are, according to Marris, the most important source of finance for the growth of capital. However, the firm is not free to retain as much profits as it might wish, because distributed profits must be adequate to satisfy the share-holders and avoid a fall in the price of shares which would render the firm attractive to take-over raiders. If distributed profits are low the existing shareholders may decide to replace the top management. If the low profits lead to a fall in the price of shares, a take-over raid may be successful and the position of managers is thus endangered.

The three financial ratios are combined (subjectively by the managers) into a single parameter $\bar{a}$ which is called the 'financial security constraint'.

This is exogenously deter-mined, by the risk attitude of the top management. Marris does not explain the process by which $\bar{a}$ is determined. It is stated that it is not a simple average of the three ratios, but rather a weighted average, the weights depending on the subjective decisions of managers.

Two points should be stressed regarding the overall financial constraint $\bar{a}$. First: Let

$$a_1 = \text{liquidity ratio} = \frac{L}{A}$$

$$a_2 = \text{leverage ratio} = \frac{D}{A}$$

$$a_3 = \text{retention ratio} = \frac{\Pi_R}{\Pi}$$

Marris postulates that the overall a is negatively related to a1, and positively to a2 and a3. That is, $\bar{a}$ increases if either the liquidity is reduced, or the debt ratio is raised by increasing external finance (loans), or the proportion of retained profits is increased. Similarly, $\bar{a}$ declines if the managers increase the liquidity of the firm, or reduce the pro-portion of external finance (D/A), or reduce the proportion of retained profits (that is, increase the distributed profits), or a combination of all three.

Secondly: Marris implicitly assumes that there is a negative relation between 'job security' (s) and the financial constraint $\bar{a}$: if $\bar{a}$ increases (by either reducing $a_1$ or in-creasing $a_2$ or increasing $a_3$) clearly the position of the firm becomes more vulnerable to bankruptcy and/or to take-over raids, and consequently the job security of managers is reduced. Thus, a high value of $\bar{a}$ implies that the managers are risk-takers, while a low value of $\bar{a}$ shows that managers are risk-avoiders.

The financial security constraint sets a limit to the rate of growth of the capital supply, $g_c$, in Marris model.

**III. The Model: Equilibrium of the Firm:**

The managers aim at the maximisation of their own utility, which is a function of the growth of demand for the products of the firm (given the security constraint)

$$U_{managers} = \text{f}\,(g_D)1$$

The owners-shareholders aim at the maximisation of their own utility which Marris assumes to be a function of the rate of growth of the capital supply (and not of profits, as the traditional theory postulated)

$$U_{owners} = f^*\,(g_c)$$

The firm is in equilibrium when the maximum balanced-growth rate is attained, that is, the condition for equilibrium is

$$g_D = g_c = \text{g* maximum}$$

The first stage in the solution of the model is to derive the 'demand' and 'supply' functions, that is, to determine the factors that determine $g_D$ and $g_c$ .

Marris establishes that the factors that determine $g_D$ and $g_c$ can be expressed in terms of two variables, the diversification rate, d, and the average profit margin, m.

**The Instrumental Variables:**
The firm will first determine (subjectively) its financial policy, that is, the value of the financial constraint $\bar{a}$, and subsequently it will choose the rate of diversification d, and the profit margin m, which maximise the balanced-growth rate g*.

**The following are policy variables in the Marris model:**
**Firstly**, a implies freedom of choice of the financial policy of the firm. The firm can affect its rate of growth by changing its three security ratios (leverage, liquidity, dividend policies).

**Secondly,** the firm can choose its diversification rate, d, either by a change in the style of its existing range of products, or by expanding the range of its products.

**Thirdly,** in Marris's model price is given by the oligopolistic structure of the industry.

Hence price is not actually a policy variable of the firm. The determination of the price in the market is very briefly mentioned in Marris's article. He argues that eventually a price structure will develop in which the market shares are stabilized. This equilibrium will be arrived at either by tacit collusion, or after a period of war during which price competition, advertising, product variation or all three weapons are used.

The length of time involved and the level of price and the number of firms which will remain in business is uncertain, due to 'imperfect knowledge of the competitors' strength, determination, and skill", and from the unpredictability of games containing chance moves.

From this line of argument, it seems that Marris is not concerned with price determination in oligo-polistic markets, but rather takes it for granted that a price structure will eventually develop. Thus, Marris seems to treat price as a parameter (given) rather than as a policy variable at the discretion of the firm. Similarly, Marris assumes that production costs are given.

**Fourthly,** the firm can choose the level of its advertising. A, and of its research and development activities, R&D. Since the price, P, and the production costs, C, are given, then it is obvious that a higher A and/or R&D expenditures will imply a lower average profit margin and, vice

versa, a low level of A and/or R&D implies a higher average profit rate. Implicit in Marris's model is the average-cost pricing rule

$$\overline{P} = \overline{C} + A + (R \& D) + m$$

where $\overline{P}$ = price, given from the market

$\overline{C}$ = production costs, assumed given

A = advertising and other selling expenses

R & D = research and development expenses

m= average profit margin

Clearly m is the residual

$m = \overline{P} - \overline{C} - (A) - (R \& D)$

Given $\overline{P}$ and $\overline{C}$, m is negatively correlated with the level of advertising and R & D ex-penditures. Thus,m is used as a proxy for the policy variables A and R&D.

In summary, all the policy variables are combined into three instruments:

$\overline{a}$, the financial security coefficient

d, the rate of diversification

m, the average profit margin

The next step is to define the variables that determine the rate of growth of demand, $g_D$, and the rate of growth of supply,$g_C$ , and express these rates in terms of the policy variables, $\overline{a}$, d and m.

**The rate of growth of the demand: $g_D$**

It is assumed that the firm grows by diversification. Growth by merger or take-over is excluded from this model. The rate of growth of demand for the products of the firm depends on the diversific-ation rate, d, and the percentage of successful new products, k, that is,

$$g_D = f_1(d, k)$$

where d = the diversification rate, defined as the number of new products introduced per time period, and k = the proportion of successful new products.

**Diversification May Take Two Forms:**

**Firstly,** the firm may introduce a completely new product, which has no close substitutes, which creates new demand and thus competes with other products for the income of the consumer. (Marris seems to narrow his analysis to firms producing consumers' goods.) This Marris calls differentiated diversification, and is considered the most important form in which the firm seeks to grow, since there is no danger of encroaching on the market of competitors and hence provoking retaliation.

**Secondly,** the firm may introduce a product which is a substitute for similar commodities already produced by existing competitors. This is called imitative diversification, and is almost certain to induce competitors' reactions. Given the uncertainty regarding the reactions of competitors the firm prefers to diversify with new products. The greater d, the higher the rate of growth of demand.

The proportion of successful new products, k, depends on the rate of diversification d, on their price, the advertising expenses, and the R & D expenditure, as well as on the intrinsic value of the products

$$k = f_3 \text{ (d, P, A, R\&D, intrinsic value)}$$

Regarding the intrinsic value of the new product Marris seems to adopt Galbraith'sand Penrose's thesis (rather far-fetched) that a firm can sell almost anything to the consumers by an appropriately organised selling campaign, even against consumers' resistance. He implicitly combines intrinsic value with price, that is, price is associated with a given intrinsic value. Price is assumed to have reached equilibrium in some way or another. Thus, price is taken as given, despite the fact that the product is new.

k depends on the advertising, A, the R & D expenditures and on d. The higher A and/or R&D, the higher the proportion of successful new products and vice versa. Marris uses m, the average profit margin as a proxy for these two policy variables. Given that m is negatively related to A and R&D, the proportion of successful new products is also negatively correlated with the average profit margin.

Finally, k depends on d, the rate of new products introduced in each period if too many new products are introduced too fast, the proportion of fails increases. Thus, although the rate of growth of demand, $g_D$ , is positively correlated with the diversific-ation rate (d) $g_D$, increases at a decreasing rate as d increases, due to the rate of intro-duction of new products outrunning the capacity of the personnel involved in the devel-opment and the marketing of the products.

There is an optimal rate of flow of 'new ideas' from the R & D department of the firm. If the research team is pressed to speed up the development process of new products there is no time to 'research' the product and/or its marketability adequately. Furthermore, top management becomes overworked when the rate of introduction of new products is high, and the proportion of unsuccessful products is bound to increase.

In summary

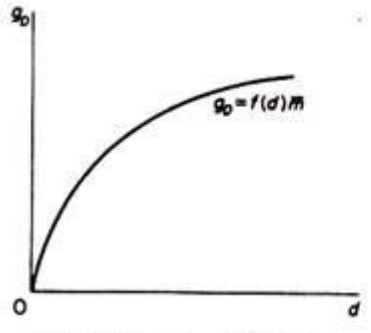
$$g_D = f_1(d, m)$$

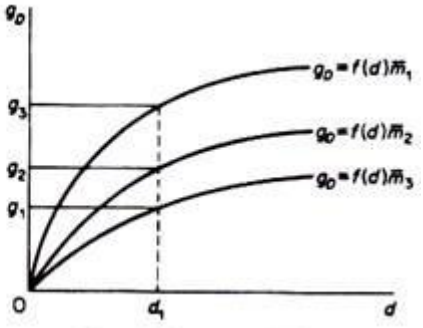
$$\frac{\partial g_D}{\partial d} > 0 \text{ (but declining)}$$

$$\frac{\partial g_D}{\partial m} < 0$$

Mathematical Expression for Optimal Rate of Flow of 'New Ideas'
The $g_D$ function is shown in following diagram

**Fig.No. 7.1** **Fig.No. 7.2**




**$g_d$= f($d$) given m** **$g_d$= f($d$, $m$) given m**

**Optimal Rate of Flow of 'New Ideas':**
The average rate of profit is constant along any gD curve. But the curve shifts down-wards as m increases ($\bar{m_1} < \bar{m_2} < \bar{m_3}$). This is due to the negative relationship between $g_D$ and m. With a given rate of diversification (for example, at $d_1$ in above figure) and given the price of the products, the lower m, the larger the A and/or the R & D expenses, and hence the larger the proportion of successful products and the higher the growth of demand ($g_3 > g_2 > g_1$). Of course, the monotonic positive relationship between d and A (and R & D), which is implied by Galbraith's and Penrose's hypothesis and is adopted by Marris, is highly questionable on a priori and empirical grounds.

**The Rate of Growth of Capital Supply:**
It is assumed that the shareholder-owners aim at the maximisation of the rate of growth of the corporate capital, which is taken as a measure of the size of the firm. Corporate capital is defined as the sum of fixed assets, inventories, short-term assets and cash reserves. It is not stated why the shareholders prefer growth to profits in periods during which growth is not steady.

The rate of growth is financed from internal and external sources. The source of in-ternal finance for growth is profits. External finance may be obtained by the issue of new bonds or from bank loans. The optimal relation between external and internal finance is still strongly disputed in economic literature.

Marris takes the position that the main source of finance for growth is profits, on the following grounds. Firstly, the issue of new shares as a means of obtaining funds is, for prestige and other reasons, not often used by an established firm. Secondly, external finance is limited by the security attitude of managers, that is, from their desire to avoid mass dismissal. Financial security is achieved by setting an upper limit to the debt/assets ratio (leverage) and a lower limit to the liquidity ratio in the long run.

Although profits are the main source of finance for growth, the top management can-not retain as much profits as it would like. There is an upper limit to the 'retention ratio', set by the desire of managers to distribute a satisfactory dividend, which will keep shareholders happy and avoid a fall in the prices of shares. Otherwise the selling of shares, or a successful take-over raid, would endanger the position of managers.

The three security ratios are subjectively determined by the managers through the security parameter a, which is a determinant of the retained profits, and hence a deter-minant of the rate of growth of capital.

Under Marris's assumptions the rate of growth of capital supply is proportional to the level of profits

$$g_c = \bar{a}(\Pi)$$

where $\bar{a}$ = the financial security coefficient
$\Pi$ = level of total profits

The security coefficient a is assumed constant and exogenously determined in this model. This assumption is relaxed at a later stage. It should be stressed, however, that so long as a is constant, growth, gc, and profits, $\Pi$, are not competing goals, but are positively related higher profits imply higher rate of growth.

The next step is to express gc in terms of the policy variables d and m. The level of total profits depends on the average rate of profit, m, and on the efficiency of the performance of the firm as reflected by its overall capital output ratio, K/X:

$$\Pi = f_4 \text{ (m, K/X)}$$

It is intuitively obvious that n and m are positively correlated (an increase in the aver-age profit margin results in an increase in the total profits)

$$\partial\Pi/\partial m > 0$$

The relationship between $\Pi$ and the capital/output ratio is more complicated. The capital/output ratio is claimed to be a measure of efficiency of the activity of the firm, given its human and capital resources. The overall K/X ratio is not a simple arithmetic average of the capital/output ratios of the individual products of the firm, but is a func-tion of the diversification rate d

$$(K/X) = f_5(d)$$

Given K, the relation between X and d is up to a certain level of d positive, reaches a maximum, and subsequently output declines with further increases in the number of new products the overall output increases initially with d due to a better utilization of the team in the R & D department as well as of the skills of the existing managerial team.

Output reaches a maximum when the d is at its optimum level allowing the optimal use of the managerial team and the R & D personnel. Beyond that point, the total output X decreases with further increases in d, and the efficiency of the firm falls the R&D per-sonnel are overworked and the decision-making process becomes inefficient, as there is not enough time allowed for the development of new products or for the study of their marketability. Hence the success rate for new products falls and efficiency declines.

Substituting for K/X in the profit function we obtain $\Pi = f_4(m, d)$

The relationship between n and d is initially positive, reaches a maximum, and then declines as d is further accelerated.

We next substitute Π in the gc function $g_c = a.[f_4(m, d)]$

The rate of growth of capital is determined by three factors the financial policies of the managers, the average rate of profit and the diversification rate.

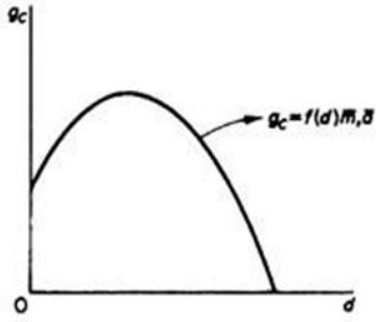
Marris assumes in his initial model that a is a constant parameter exogenously deter-mined by the risk-attitude of managers, while there is a positive relation between gc and m

$$\partial g_C / \partial m > 0$$

The relationship between gc and d is not monotonic. The rate of growth of capital, gc, is positively correlated with d up to the point of optimal use of the R & D personnel and the team of managers; but gc is negatively correlated with d beyond that point a higher d implies hastening up of the diversification process → inefficient decisions → fall in the overall profit level → low availability of internal finance and consequently a lower rate of growth gc.

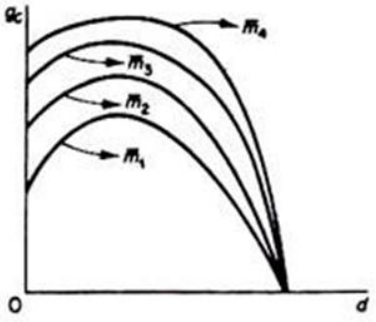
The relation between gc and d, keeping a and m constant, is shown in figure 7.3. If we allow both d and m to change, while keeping a constant, we obtain a family of gc = f2 (d, m) curves (figure 7.4). The average profit rate is depicted as a shift factor of the g c = f(0) curve. The higher the average profit rate, the further from the origin the gc curves will be (m1 < m2 < m3). These curves are drawn under the assumption that a is constant. (The effects of a change in a are discussed in section IV below.)

**Figure No. 7.3**




**$g_d$= f(*d*) given m**

**Figure No. 7.4**




**$g_c$= f(*d, m*) given m**

Summarising the above arguments, we may present Marris's model in its complete form as follows:

gD = f1(m, d) – (demand-growth equation)

Π = f4(m, d) – (profit equation)

gC = a.[f4(m, d)] – (supply-of-capital equation)

a < a* (security constraint)

gD = gc (balanced-growth equilibrium condition)

a is exogenously determined by the risk-attitude of managers. The level of profit Π is endogenously determined. The variables m and d are the policy instruments. Given the balanced-growth equilibrium condition, we have in fact one equation in two unknowns (m and d, given a)

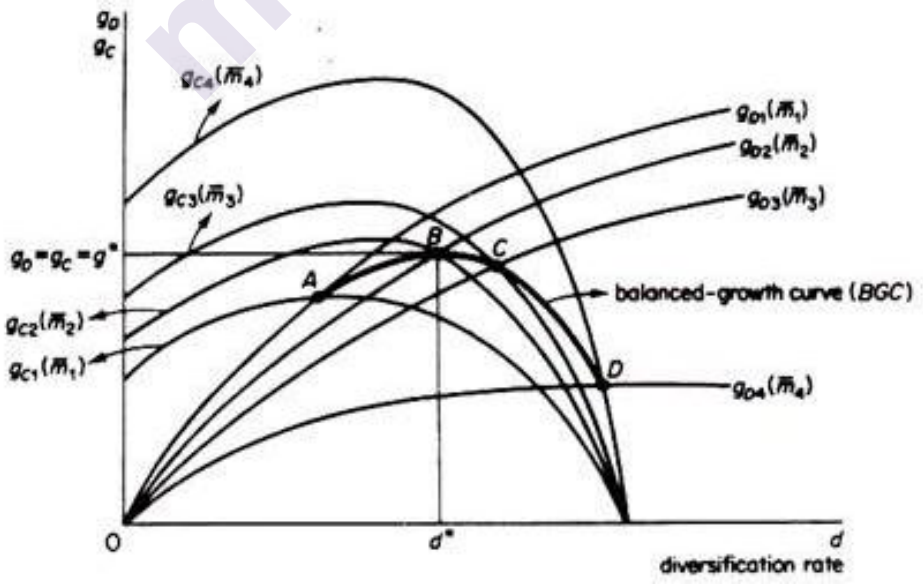
$$f1(m, d) = a.[f4(m, d)]$$

**Equilibrium of the firm:**

Clearly the model cannot be solved (is under identified), unless one of the variables m or d is subjectively determined by the managers. Once the managers define a and one of the other two policy variables, the equilibrium rate of growth can be determined.

The equilibrium of the firm is presented graphically in figure 16.5, formed by super-imposing figures 16.2 and 16.4. Given their shapes, the gD and gc curves associated with a given profit rate intersect at some point. For example, the gD and gc curves corres-ponding to m, intersect at point A; the gD and gc curves associated with m2 intersect at point B, and so on. If we join all points of intersection of gD and gc curves corresponding to the same level of m we form what Marris calls the balanced-growth curve (BGC), given the financial coefficient a.

**Figure No. 7.5**



**Equilibrium of Firm in Marris Model:**

The firm is in equilibrium when it reaches the highest point on the balanced-growth curve. The firm decides its financial policy, denoted by a. It next chooses subjectively a value for either m or d. With these decisions taken, the firm can find its maximum bal-anced-growth rate, consistent with a and with the chosen value of one of the other two policy variables. In figure 7.5 the BGC corresponding to a is ABCD.

The balanced- growth rate g* is defined by the highest point B of this BGC. This g* rate is compatible with a unique pair of values of the policy variables, m* and d*. If the firm chooses d*, then m* is simultaneously determined; alternatively, if the firm chooses m*, then d* is simul-taneously determined from the function

$g^* = f1\ (m^*, d^*) = a.\ [f4\ (m^*, d^*)]$

Substituting m* and d* in the profit function

$$\Pi = a[f4\ (m, d)]$$

we find the level of profit, Π*, required to finance the balanced-growth rate, g*. Thus profit is endogenously determined in Marris's model. Furthermore, growth and profit are not competing goals (so long as a is constant). From the gc function

$$gc = a\ .\ (\Pi)$$

it is obvious that higher profit implies higher growth rate. However, if the financial coefficient a is allowed to vary, then profits and growth become competing goals.

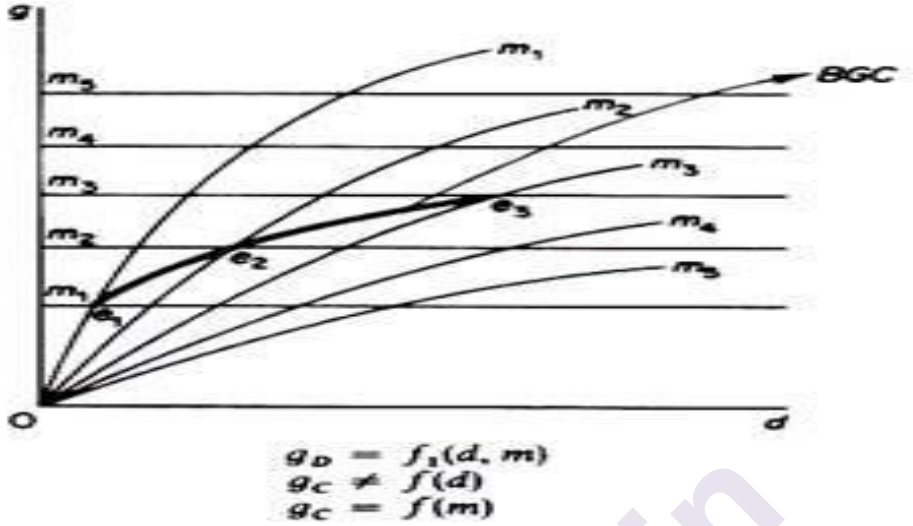
The question is does the BGC have a maximum? Marris argues that so long as either (or both) of the gc or gD curves flattens out or bends, there will always be a maximum point on the BGC curve. Furthermore, depending on the shape of the gc and the gD curves, the BGC may be platykurtic, that is, have a flat stretch which indicates that there are several optimal solutions the g* may be achieved by a large number of com-binations of the values of the policy variables m and d (given a is already chosen).

It is only if the gc curve is parallel to the 4-axis ($gc = f(m)$ but $gc \neq f(d)$) and the gD curves are straight upwards-sloping curves (implying that $gD = f(d, m)$, but $k \neq f(d)$ and hence the gD curve does not flatten out) that the BGC increases continuously, never reaching a maximum. This situation is, however, improbable given the capacity for efficient de-cision making of the managerial team and the capacity for well-explored new products of the R & D department of the firm.

These cases are graphically shown in figures 7.6-7.9. Figure 7.6 depicts the case where $gc \neq (d)$, while $gD = f(d, m)$. The gc curve becomes parallel to the d-axis, show-ing that gc does not vary as d increases. The gc curve shifts upwards (parallel to itself) as the average profit margin

increases, given that gc and m are positively related. The balanced-growth curve has a maximum defined by the curvature of the gD = f (m, d) function (the maximum g occurs at point e3 in figure 16.6).

**Figure No. 7.6**

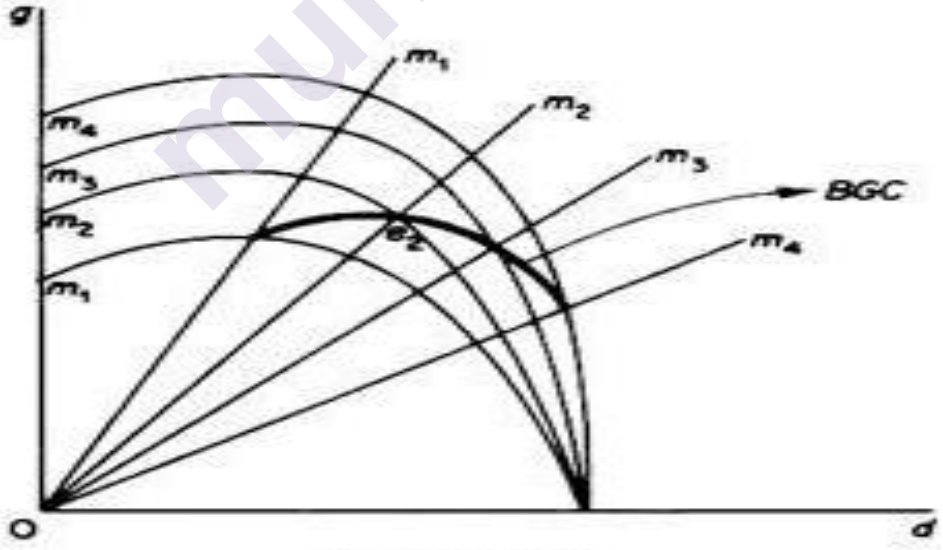



$g_D = f_1(d, m)$
$g_C \neq f(d)$
$g_C = f(m)$

**Balanced-Growth Curve of Marris Model:**

Figure 7.7 depicts the case where gn = *f*x(m, d), and gc = *f*2(d, m). But the curve gD becomes a straight line through the origin, showing that gD has a constant slope irrespective of changes in the diversification rate. The gD curve (line) shifts downwards towards the x-axis as m increases. The balanced-growth curve has still a maximum (e2) due to the curvature of the gc function.

**Figure No. 7.7**

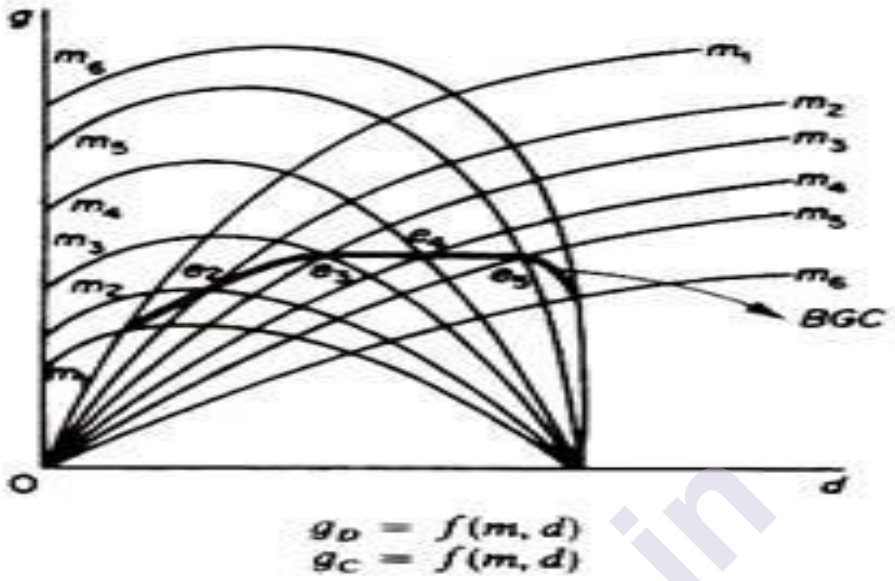



$g_D = f(d, m)$
(with $\partial g_D/\partial d$ constant)
$g_C = f(d, m)$

Figure 7.8 shows a platykurtic balanced-growth curve the gD and gc functions have several points of intersection (due to their shapes) that lie

on a straight line. The flat part of the balanced-growth curve implies that the same optimal (maximum) g* may be achieved by a very large number of combinations of m and d.

**Figure No. 7.8**

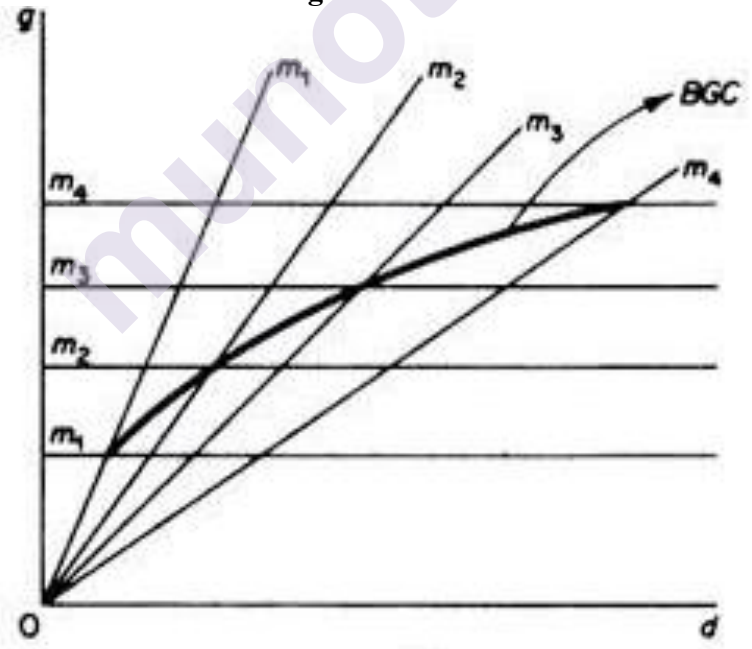



$$g_D = f(m, d)$$
$$g_C = f(m, d)$$

**Platykurticf Balanced-Growth Curve in Marris Model:**

Finally figure 16.9 shows the improbable case a balanced-growth curve which never reaches a maximum (explosive growth).

**Figure No. 7.9**




$$g_D = f(m, d), \quad \frac{\partial g_D}{\partial d} \text{ constant}$$
$$g_D \neq f(d), \quad g_C = f(m)$$

**Improbable Case of Balanced-Growth Curve in Marris Model**

**IV. Maximum Rate of Growth and Profits:**

Marris argues that in the real world the financial coefficient a is not a constant, but varies. Changes in a clearly affect gc, given

$$gc = a(\Pi) = a\,[f4(m, d)]$$

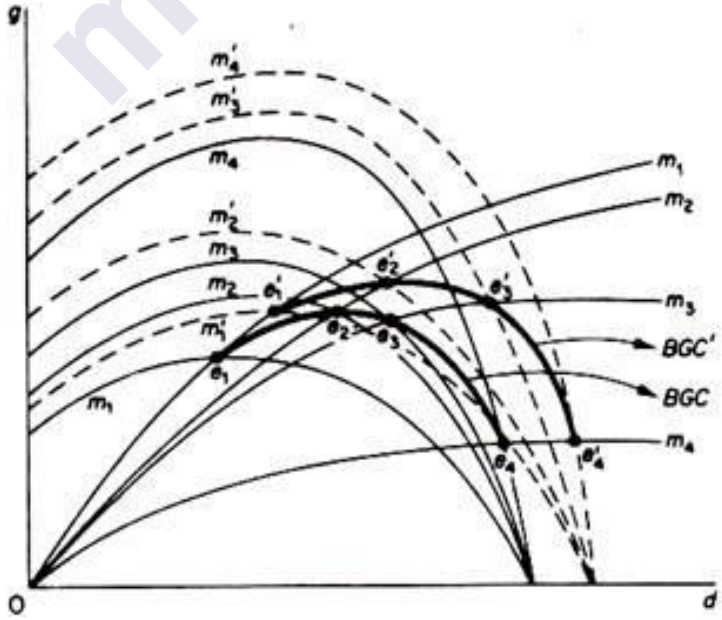
A change in a will shift the gc curves if a increases the gc curves will shift upwards, while if a is reduced the gc curves will shift downwards. The new set of gc intersects the given set of gD curves at new points, which form a new balanced-growth curve. Given that the relationship between gc and a is positive ($\partial gD/\partial a > 0$), an increase in a leads to an increase in the rate of growth.

An increase in a will occur if one or more of the three security ratios changes as follows a is higher if the liquidity ratio (a1) is lowered; or if the debt ratio (a2) is increased; or if the retention ratio (a3) is increased. This is due to the fact that a is positively related to a2 and a3, but negatively related to ay.

Clearly an increase in a, however realised, implies a less 'prudent', more risky policy of the managers, since a decrease in the liquidity ratio, or an increase in the indebtedness or an increase in the retained profits (which implies a reduction in the paid dividends) reduces the job security of the managers.

Graphically an increase in a is shown by an upwards shift of the BGC (to the position A'B'C'D' in figure 16.10). Given the gD curves, the highest point of the new BGC will be above the highest point of the original BGC. This implies that the balanced rate of growth g cannot be maximised unless a assumes its highest optimal value a*. Con-sequently in equilibrium a – a*, that is, the financial constraint takes the form of equality at equilibrium.

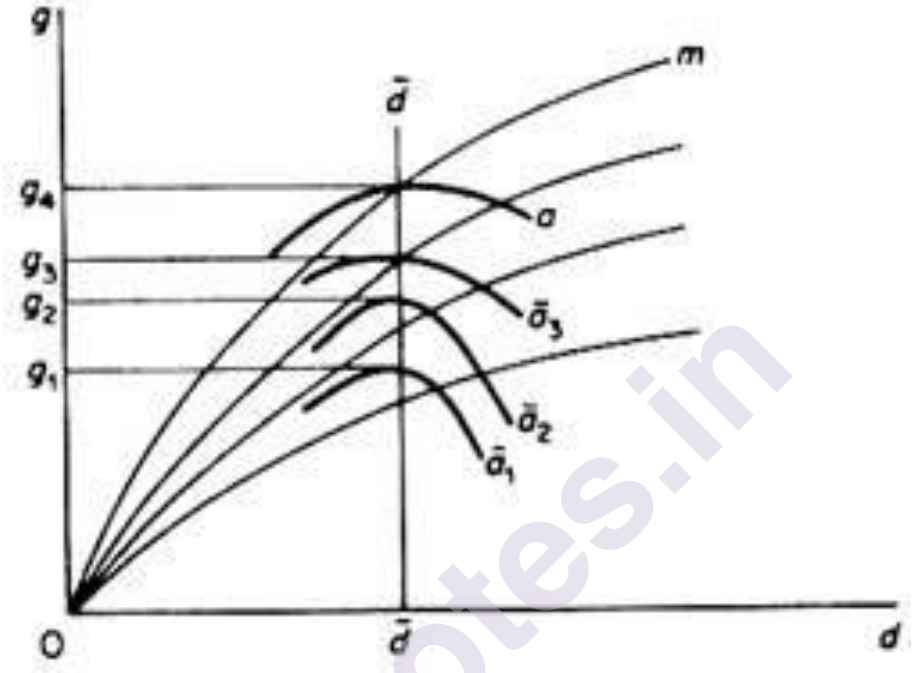
**Figure No. 7.10**




**Maximum Rate of Growth and Profit in Marris Model**

Marris next argues that if a is allowed to vary, growth and profits may become com-peting goals. If a is lowered below its optimum value a* the growth rate is reduced but the profit level, Π, may be raised. A lower value of a (given the d rate) denotes a shift to a lower balanced-growth curve, which implies the intersection of gD and gc curves cor-responding to a higher m, and hence a higher n, since n is a positive function of m (figure 7.11).

**Figure No. 7.11**




**Profit Maximisation in Marris Model:**

Thus, although when a is held constant, maximising the growth rate implies maximising profit (g and Π are not competing goals), when a is allowed to vary, growth and profits become competing goals if a is treated as a variable, the firm cannot maxi-mise both the rate of growth and profit.

This explains that under some circumstances managers' objectives (for higher g) and stockholders' objectives (for higher Π) may con-flict. It should, however, be clear that a cannot be increased beyond a certain value, determined by the minimum profit requirements of the shareholders; otherwise the job security of managers decreases dangerously.

If the solution of the model does not yield in adequate to satisfy the stockholders, a will be reduced (via, for example, a lowering of the retention ratio), until the maximum obtainable balanced-growth rate is consistent with a level of profit that is satisfactory. This implies that managers seek to maximise the growth rate subject to a minimum profit constraint.

## 7.3 WILLIAMSON'S MODEL OF MANAGERIAL DISCRETION

Oliver E. Williamson hypothesised (1964) that profit maximization would not be the objective of the managers of a joint stock organisation. This theory, like other managerial theories of the firm, assumes that utility maximisation is a manager's sole objective. However it is only in a corporate form of business organisation that a self-interest seeking manager maximise his/her own utility, since there exists a separation of ownership and control. The managers can use their 'discretion' to frame and execute policies which would maximise their own utilities rather than maximising the shareholders' utilities. This is essentially the principal–agent problem. This could however threaten their job security, if a minimum level of profit is not attained by the firm to distribute among the shareholders.

The managerial theory of firm developed by Oliver E. Williamson states that managers apply discretion in making and implementing policies to maximize their own utility rather than trying for the maximization of profit which ultimately maximize own utility subject to minimum profit. Profit works as a limit to the top managers' behaviour in the sense that the financial market and the shareholders require a minimum profit to be paid out in the form of dividends, otherwise the job security of managers is put in danger. Hence, managers look at their self-interest while making decision on price and selling quantity of output. Manager's decision on price and output differs from the decisions of profit maximizing firm.

Utility maximization of managers guided by their own self-interest is possible, like in Baumol's sales maximization model, only in a corporate type of business organization with the separation of ownership and management functions. Such organizational structure permits the managers of a firm to pursue their own self-interest, subject only to their ability to keep effective control over the firm. In particular managers are fairly certain of keeping hold of their power (i) if profits at any time are at an acceptable level, (ii) if the firm shows a reasonable rate of growth over time, and (iii) if sufficient dividends are paid to keep the stockholders happy.

Williamson's model suggests that manager's self-interest focuses on the achievement of goals in four particular areas, namely:

1. High salaries
2. Staff under their control
3. Discretionary investment expenditures
4. Fringe benefits (i.e., additional employee benefit: an additional benefit provided to an employee, for example, a company car or health insurance)

**The basic assumptions of the model are:**

1. Imperfect competition in the markets.
2. Weakly competitive environment.
3. Divorce of ownership and management. A divorce of ownership from control of firm (manager is free to perform any action)
4. A minimum profit constraint exists for the firms to be able to pay dividends to their share holders.A capital market imposes minimum profit constraint (manager's work for minimum profit imposed by a capital market).

**Managerial Utility Function:**

The managerial utility function includes variables such as salary, job security, power, status, dominance, prestige and professional excellence of managers. Of these, salary is the only quantitative variable and thus measurable. The other variables are non-pecuniary, which are non-quantifiable. The variables expenditure on staff salary, management slack, discretionary investments can be assigned nominal values. Thus, these will be used as proxy variables to measure the real or unquantifiable concepts like job security, power, status, dominance, prestige and professional excellence of managers, appearing in the managerial utility function.

**Utility function or "expense preference"** of a manager can be given by:

$$U = U(S, M, I_D)$$

where U denotes the Utility function, S denotes the "monetary expenditure on the staff", M stands for "Management Slack" and ID stands for amount of "Discretionary Investment".

**"Monetary expenditure on staff"** include not only the manager's salary and other forms of monetary compensation received by him from the business firm but also the number of staff under the control of the manager as there is a close positive relationship between the number of staff and the manager's salary.

**"Management slack"** consists of those non-essential management perquisites such as entertainment expenses, lavishly furnished offices, luxurious cars, large expense accounts, etc. which are above minimum to retain the managers in the firm. These perks, even if not provided would not make the manager quit his job, but these are incentives which enhance their prestige and status in the organisation in turn contributing to efficiency of the firm's operations. The Management Slack is also a part of the cost of production of the firm.

**"Discretionary investment"** refers to the amount of resources left at a manager's disposal, to be able to spend at his own discretion. For example, spending on latest equipment, furniture, decoration material, etc. It satisfies their ego and gives them a sense of pride. These give a boost to the manager's esteem and status in the organisation. Such investments are

over and above the amount required for the survival of the firm (such as periodic replacement of the capital equipment).

**Concepts of profit in the model:**
The various concepts of profit used in the model needs to be understood clearly before moving to the main model. Williamson has put forth four main concepts of profits in his model:
**Actual profit (Π):**

$$\Pi = R - C - S$$

where R is the total revenue, C is the cost of production and S is the staff expenditure.

**Reported profit (Πr)**

$$\Pi_r = \Pi - M$$

where Π is the actual profit and M is the management slack.

**Minimum Profit ($\Pi_0$):**
It is the amount of profit after tax deducted which should be paid to the shareholders of the firm, in the form of dividends, to keep them satisfied. If the minimum level of profit cannot be given out to the shareholders, they might resort of bulk sale of their shares which will transfer the ownership to other hands leaving the company in the risk of a complete take over. Since the shareholders have the voting rights, they might also vote for the change of the top level of management. Thus the job security of the manager is also threatened. Ideally the reported profits must be either equal to or greater than the minimum profits plus the taxes, as it is only after paying out the minimum profit that the additional profit can be used to increase the managerial utility further.

$$\Pi_r \geq \Pi_0 + T$$

where Πr is the reported profit, Π0 is the minimum profit and T is the tax.

**Discretionary profit (ΠD):**
It is basically the entire amount of profit left after minimum profits and tax which is used to increase the manager's utility, that is, to pay out managerial emoluments as well as allow them to make discretionary investments.

$$\Pi_D = \Pi - \Pi_0 - T$$

where ΠD is the discretionary profit, Π is the actual profit, Π0 is the minimum profit and T is the tax amount.

However, what appears in the managerial utility function is discretionary investments (ID) and not discretionary profits. Thus it is very important to distinguish between the two as further in the model we would have to maximize the managerial utility function given the profit constraint.

$$I_D = \Pi_r - \Pi_0 - T$$

where Πr is the reported profit, Π0 is the minimum profit and T is the tax amount.

Thus it can be seen that the difference in the Discretionary Profit and the Discretionary investment arises because of the amount of managerial slack. This can be represented by the given equation

$$\Pi_D = I_D + M$$

where ΠD is the discretionary profit, ID is the Discretionary investment and M is the management slack.

**Model Framework:**
For simple representation of the model the managerial slack is considered to be zero. Thus there is no difference between the actual profit and reported profit, which implies that the discretionary profit is equal to the discretionary investment. I.e.

$$\Pi_r = \Pi \text{ or } \Pi_D = I_D$$

where Πr is the reported profit, Π is the actual profit, ΠD is the discretionary profit and ID is the discretionary investment. Such that the utility function of the manager becomes

$$U = U(S, I_D)$$

where S is the staff expenditure and ID is the discretionary investment.

There is a trade off between these two variables. Increase in either will give the manager a higher level of satisfaction. At any point of time the amount of both these variables combined is the same, therefore an increase in one would automatically require a decrease in the other. The manager therefore has to make a choice of the correct combination of these two variables to attain a certain level of desired utility.

Substituting $I_D = \Pi - \Pi_0 - T$ in the new managerial utility function, it can be rewritten as $U = U(S, \Pi - \Pi_0 - T)$

The relationship between the two variables in the manager's utility function is determined by the profit function. Profit of a firm is dependent on the demand and cost conditions. Given the cost conditions the demand is dependent of the price, staff expenditures and the market condition.

$$X = f(S, \bar{P}, \bar{E})$$

Price and market condition is assumed to be given exogenously at equilibrium. Thus the profit of the firm becomes dependent on the staff expenditure which can be written as

$$\Pi = f(X) = f(S, \bar{P}, \bar{E})$$

Discretionary profit can be rewritten as

$$\Pi_D = f(S, \bar{P}, \bar{E}) - \Pi_0 - T)$$

In the model, the managers would try to maximise their utility given the profit constraint

$$\max U = U(S, \Pi - \Pi_0 - T)$$

$$\text{subject to} \Pi \geq \Pi_0 + T$$

**Graphical representation of the model:**

**Fig No. 7.12**
**Utility indifference curves of managers**

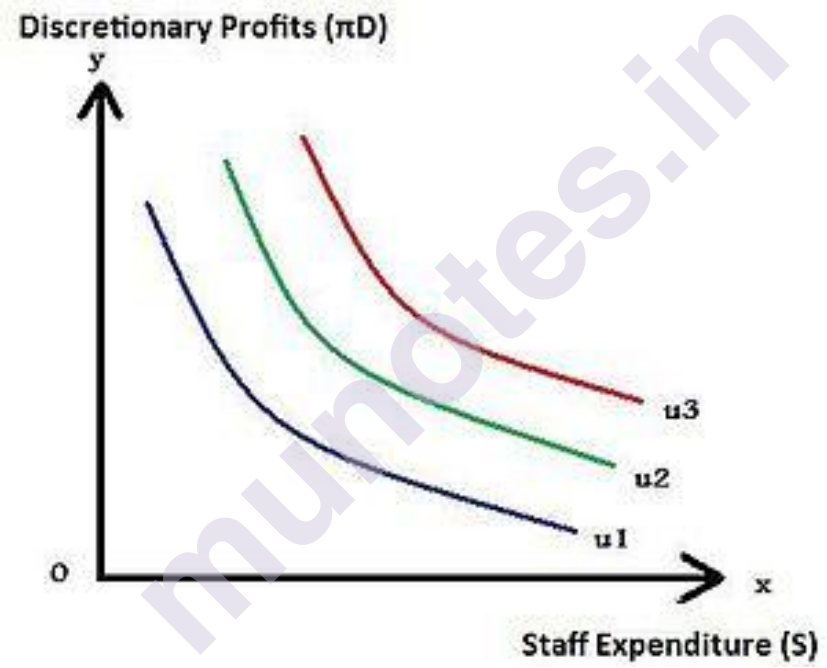
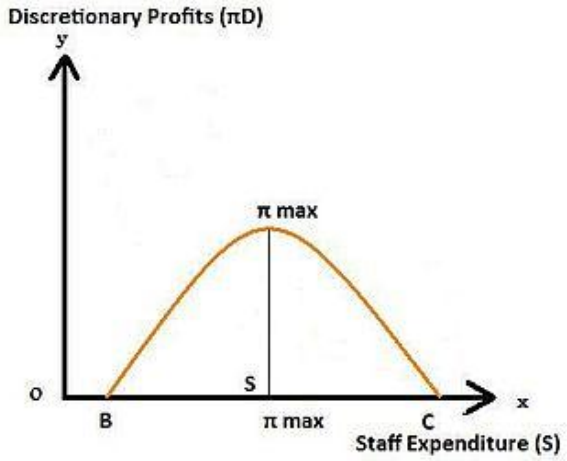



Fig 7.12. shows the various levels of utility (U1, U2, U3) derived by the manager by combining different amounts of discretionary profits and staff expenditure. Higher the indifference curve, higher is the level of utility derived by the manager. Hence the manager would try to be on the highest level of indifference curve possible given the constraints. Staff expenditure is plotted on the x-axis and discretionary profits on the y-axis. The discretionary profit in this simplified model is equal to the discretionary investment. The indifference curves are downward sloping and convex to the origin. This shows diminishing marginal rate of substitution of staff expenditure for discretionary profits. The curves are asymptotic in nature which implies that at any point of time and under any given circumstance the manager will choose positive amounts of both discretionary profits and staff expenditure.

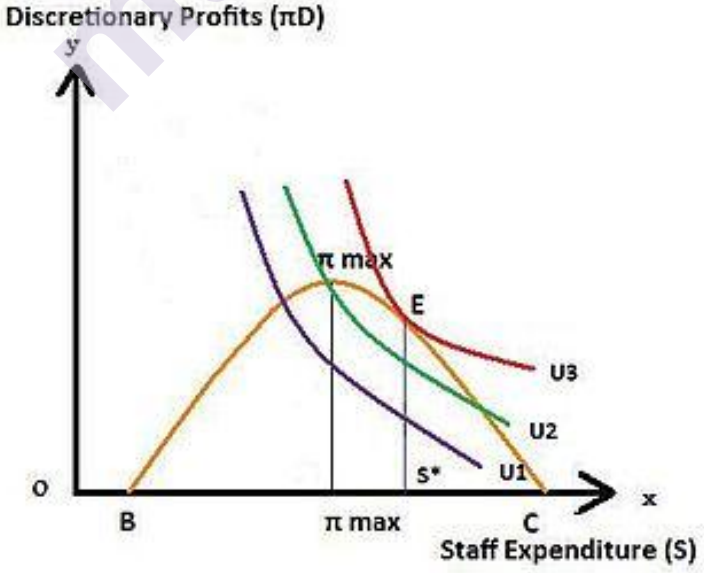
**Fig No. 7.13**
**Discretionary Profit Curve**




Assuming that the firm is producing an optimum level of output and the market environment is given, the discretionary profits curve is generated, shown in Fig 7.13. It gives the relationship between staff expenditure and discretionary profits.

It can be seen from the figure that profit will be positive in the region between the points B and C. Initially with increase in profits, the staff expenditure the discretionary profits also increase, but this is only till the point Πmax, that is, till S level of staff expenditure. Beyond this if staff expenditure is increased due to increase in output, then a fall in the discretionary profits is noticed. Staff expenditure of less than B and more than C is not feasible as it wouldn't satisfy the minimum profit constraint and would in turn threaten the job security of managers.

**Fig No. 7.14**
**Equilibrium of a firm in Williamson's Model**



To find the equilibrium in the model, Fig 7.12. is superimposed on Fig 7.13. The equilibrium point is the point where the discretionary profit curve is tangent to the highest possible indifference curve of the manager, which is point E in Fig 3. Staying at the highest profit point would require the manager to be at a lower indifference curve U2. In this case the highest attainable level of utility is U3. At equilibrium, the level of profits would be lower but staff expenditure S* is higher than the staff expenditure made at the maximum profit point. As indifference curve is downward sloping, the equilibrium point would always be on the right of the maximum profit point. Thus, the model shows the higher preference of managers for staff expenditure as compared to the discretionary investments.

**Criticism**

1. The model fails to describe how businesses take their price and output decisions in a highly competitive set up.
2. The relationship between better performance of managers and the increasing amounts spent on manager's utility by the firm is not always true.
3. The model does not apply in a dynamic set up like changing demand and cost conditions during booms and recessions.
4. This model fails to deal with the core problem of oligopolistic interdependence and of strong oligopolistic rivalry.
5. This model is applicable in markets where rivalry is not strong (for example ,in an oligopolistic market where there is some form of collusion), or for firms who have some advantage over their rivals (for example, Patent, superior know-how). However, in the long run such advantages which shelter a firm from competition are usually weakened, and competition is enhanced.

Williamson like other managerial theory of the firm assumes that utility maximization is the sole objective of the managers of a joint stock organization. It is also known as "Managerial discretion Theory". Williamson emphasize that managers are motivated by their own self-interest and they tries to maximize their own utility function. Alike Baumol sales maximization model, the utility maximization objective of the managers are subject to the constraint that after tax profits are large enough to pay dividends to the shareholders. However, it is pointed out that utility maximization by the self-interest seeking managers is possible only in corporate form of the business organization as there exists separation of ownership and control.

## 7.4 BEHAVIOURAL THEORIES OF THE FIRM

**Introduction:** The **behavioral theory of the firm** first appeared in the 1963 book A Behavioral Theory of the Firm by **Richard M. Cyert and James G. March.** The work on the behavioral theory started in 1952 when March, a political scientist, joined Carnegie Mellon University,

where Cyert was an economist. Before this model was formed, the existing theory of the firm had two main assumptions: profit maximization and perfect knowledge. Cyert and March questioned these two critical assumptions.

A behavioral model of rational choice by Herbert A. Simon paved the way for the behavioral model. Neo-classical economists assumed that firms enjoyed perfect information. In addition, the firm maximized profits and did not suffer from internal resource allocation problems.

Advocates of the behavioral approach also challenged the omission of the element of uncertainty from the conventional theory. The behavioral model, like the managerial models of Oliver E. Williamson and Robin Marris, considers a large corporate business firm in which the ownership is separate from the management.

In classical economics, the theory of firms is based on the assumption that they will seek profit maximisation. However, in the real-world managers and owners may behave quite differently. Behavioural Theories of the Firm include:

- **Size of a firm/prestige**. Some managers may simply aim for working in a big and seemingly successful firm which gives more prestige and honour. Managers may be motivated to prove their projects are successful. This can cause firms to pursue goals which have a high profile. It may explain why firms persist with projects which may not be desirable. There is a cost to letting go of past decisions.
- **Profit satisficing**. Based on the problem of asymmetric information. Owners wish to maximise profits, but, workers don't. Because owners don't have perfect information, workers and managers are able to get away with decisions that don't maximise profits.
- **Co-operative/ethical concerns**. Some firms may be set up with very different objectives to the traditional model of profit maximisation. In co-operative firms, the goal is to maximise the welfare of all stakeholders. In this model, ideas of altruism, concern for the environment and workers welfare may explain many decisions. The firm may also be set up with specific charitable aims.
- **Human emotion/bias**. The economic model of a rational economic man assumes that individuals seek to maximise their economic welfare with rational choice. However, in the real world, we are influenced by human emotion. This could be discrimination based on bias and prejudice. Or it could be irrational exuberance and the perceived wisdom of following the crowd. For example, in asset bubbles, mortgage companies can get caught up in relaxing their lending criteria and lending mortgages to those at risk of default.

**The Cyert and March Theory of Firm:**

**Firm depends on the demand of the members of the coalition:**
The behavioural theory of firm was developed by Cyert and March, focuses on the decision making process of the large multi-product firm under uncertainty in an imperfect market. They deal with the large corporate managerial business in which ownership is separated. Their theory was originated from the concern about the organizational problem with the internal structure of such firms that creates the need to investigate the effect on the decision-making process in these large organizations. The internal organizational actors may well explain the difference in the reactions of firms to the same external stimuli, that to the same changes in their economic environment.

The assumptions underlying the behavioural theories about the complex nature of the firm introduces an element of realism into the theory of the firm. The firm is not treated as a single-goal, single decision unit, as in the traditional theory, but as a multi goal, multi decision organization coalition. The firm is as a coalition of different groups which are connected with its activity; in various ways, managers, workers, shareholders, customers, suppliers, bankers, tax inspectors and so on. Each group has its own set of goals or demands.

The behavioural theory recognizes explicitly that there exists a basic dichotomy in the firm, there are individual members of the coalition firm and there is the organization coalition known as 'the firm'. The consequence of the dichotomy is a conflict of goals; individuals may have different goals to those of the organization firm.

Cyert and March argue that the goals of the firm depends on the demand of the members of the coalition, while the demand of these members are determined by various factors such as aspiration of the members, their success in the past in occupying their demands, the expectations, the achievements of other groups in the same or other firms, the information available to them. The demands of the various groups of the coalition firm change continuously over time. Given the resources of the firm in any one period, not all demands, which confront the top management can be satisfied. Hence, there is a regular bargaining process between the various members of the coalition firm and inevitable conflict.

The top management has several tasks; to get the goals of the firm which are often in conflict with the demands of the various groups, to resolve the conflict between the various groups, to reconcile as far as possible the conflict in goals of the firm and of its individual groups.

There is a strong relation between demands and past achievement. Demands take the form of aspiration levels. Demands change continuously, depending on past achievement and on changes in the firm and its environment. In any one period the demands which will actually be

presented by any particular group to the top management depend on past achievement of demands previously pursued by the particular group, on the achievement of other groups in the same firm, on the achievement of similar groups in other firms, on past aspiration levels, on expectations, and on available information.

Cyert and March argue that the relationship between demands-aspirations and past achievement depends on actual and expected changes in the performance of the firm and changes in its environment: Firstly, in a 'steady situation', with no growth or dynamic changes in the environment, aspirations (demands) and past achievement tend to become equal. Secondly, in a dynamic situation with growth, aspiration levels (demands) lag behind achievement.

This time-lag is crucial to the behavioural theory. During this time lag the firm is able to accumulate 'surpluses' or 'excess-profits', which may be used as a means of resolution of the conflict in the firm and which act as a stabiliser of the firm's activity in a changing environment. Thirdly, in a period of decline of the activity of the firm, demands are larger than past achievements, because the aspiration levels of the members of the coalition adjust downwards slowly.

This process of demand and aspiration-level formation renders the behavioural theory dynamic: the aspiration levels-demands at any time t depend on the previous history of the firm, that is, on previous levels of achievement and previous aspiration levels.

The goals of the firm are set by the top management, which the main five goals of the firm are:

1. **Production Goal:** The production goal originates from the production department. The main goal of the production manager is the smooth running of the production process. Production should be distributed evenly over time, irrespective of possible seasonal fluctuations of demand, so as to avoid excess capacity and lay off of workers at some periods and over working the plant and resorting to rush recruitment of workers at other times with the consequence of higher, costs due to excess capacity and dismissal payments or too frequent breakdowns of machinery and waste of raw materials in period of 'rush' production.

2. **Inventory Goal:** The inventory goal originates mainly from the inventory department if such a department exists, or from the sales and production department. The sales department wants an adequate stock of output for the customers, while the production department needs adequate stocks of raw materials and other items necessary for a smooth flow of the output process.

3. **Sales Goal:** The sales goal and the share of the market goal originate from the sales department. The same department will also normally

set the 'sales strategy' that is decided on the advertising campaigns, the market research programs, and so on.

4. **Profit Goal:** The profit goals is set by the management so as to satisfy the demand of share holders and the expectations of bankers and other finance institutions; and also to create funds with which they can accomplish their own goals and projects, or satisfy the other goals of the firm.

5. **Share of the market goal:** While making decisions, the firms are guided by these goals. All goals must be satisfied but there is an implicit order of priority among them. The conflict among different goals may crop up.

The number of goals of the firm may be increased, but the decision making process becomes increasing complex. The efficiency of decision making decreases as the number of goals increases. The law of diminishing returns holds for managerial work as for all other types of labor.

The goals of the firm are ultimately decided by the top management through continuous bargaining between the groups of the coalition. In the process of goal formation, the top management attempts to satisfy as many as possible of the demands with which the various members of the coalitions confront it. The goals of the firm such as the goals of the individual members or particular groups of the coalition take the form of aspiration levels rather than strict maximizing constraints.

The firm in the behavioural theories seeks to satisfy, i.e., to attain a 'satisfactory' overall performance as defined by the set aspiration goals, rather than maximize profits, sales or other magnitudes. The firm is as satisfying organization rather than a maximizing entrepreneur. The top management, responsible for the coordination of the activities of the various members of the firm, wishes to attain a 'satisfactory' level of production, to attain a share of the market, to earn a 'satisfactory' level of profit, to divert a 'satisfactory' percentage of their total receipts to research and development or to advertising, to acquire a 'satisfactory' public image and so on. But it is not clear in the behavioral theories what is a satisfactory and what an unsatisfactory attainment is.

They argue that satisfying behaviour is rational given the limitations, internal and external with in which the operation of the firm is confined. They take by the form of aspiration levels, and whether attained, the performance of the firm is considered as satisfactory. The goals do not normally take the form of maximization of the relevant magnitudes. The firm is not a maximizing but rather a satisfying organization.

Some of the above goals may be desirable to (and consequently acceptable by) all members of the coalition. For example**, the sales goal** is directly desirable to the sales manager and his department, to the top management and most probably to the shareholders. But this goal is also indirectly

desirable to all the other members of the coalition, since all groups know that unless the firm sells whatever it produces no one will be able to attain his own individual goals.

Other goals are desirable to only some of the groups. For example, profits are the concern of the shareholders and the top management, but not of the employees in lower administrative levels or of the workers 'on the floor.' The conflicts arising in the process of goal-setting at the level of top management are resolved by various means which are examined below.

**Conflicting Goals:**

The aspiration levels of the individuals within the firm which determine these goals change over time as a result of organizational learning. Thus, these goals are regarded as the product of a bargaining learning process in the organization coalition. But it is not essential that the different goals may be resolved amicably. There may be conflicts among these goals.

The conflicting interest can be reconciled by the distribution of side payments' to members of the coalition. Side payments may be in cash or kind, the latter being mostly in the form of policy side payments. But the actual total side payments is not fixed for the coalition but depends upon the demand of members and on the form of the coalition. Demands of coalition members equal actual side payments only in the long-run. But the behavioral theory focuses on the short-run relation between side payments and demands and on the imperfections in factor markets.

In the short-run, new demands are being constantly made and the goals of the organization are continually adapted, to a greater or lesser extent, to take account of these demands. The demands of the members of the organizational coalition need not be mutually consistent. But all demands are not made simultaneously and the organization can remain viable by attending the demands in sequence. A problem will arise when the organization is not able to accommodate the demands of its members even sequentially, because it lacks the resources to do so.

Besides, side payments, the conflicting goals of the organization are resolved by subjecting them to a constant review. This is because, aspiration levels' of coalition members change with experience. In fact, the aspiration levels change with the process of satisfying. Each person in the organization has a satisfying level for each of his goals

**Uncertainty and the Environment of the Firm:**

Cyert and Match-distinguish two types of uncertainty: market uncertainty and uncertainty of competitors' reactions. Market uncertainty refers to possible changes in customers' preferences or changes in the techniques of production. This form of uncertainty is inherent in any market structure. It can partly be avoided by search activity and information-gathering, but it cannot be avoided completely. Given the market uncertainty the managerial firm avoids long-term planning and works within a short time-horizon. The behavioural theory postulates that the firm considers only the

short-run and chooses to ignore the long-run consequences of short-run decisions.

The uncertainty arising from competitors' actions and reactions, that is, from oligopolistic interdependence, is brushed aside by this theory by assuming that existing firms have arrived at some form of tacit collusion. The various forms of trade associations, clubs and the issue of various 'informative' bulletins or other publications provide a means by which firms give out information concerning their prices or future outlays of various kinds, expecting every other competitor to do the same.

This sort of modus vivendi is called a 'negotiated environment' by Cyert and March. The firm is assumed to 'negotiate' in some way or another with its competitors so as to avoid uncertainty. Thus, the core problem of oligopolistic markets that of competitors' interdependence, is 'solved' by assuming collusive action of the firms.

In general, the theory pays too little attention to the environment and its effect on the goal-formation process and the pricing and output decisions at the level of top management. It examines internal resource allocation, assuming collusion with competitors. It says nothing about the threat of potential entry which is crucial in the present world of mergers and continuous diversification.

The environment is taken as given and as such is practically ignored in the analysis of the behaviour of the firm. This ignoring of the environment is apparent in the model that follows, which is used by Cyert and March as an illustration of the workings of their theory. The rules by which demand and costs are estimated, the rules for investment decisions and other crucial steps in the analysis are too mechanical.

**A Simple Model of Behaviourism:**

Here we briefly present the simple model used by Cyert and March as an illustration of the decision-making process within the modern large corporation. The model refers to the case of a duopoly. The decision process involves the determination of the output which is homogeneous, so that a single price will ultimately prevail in the market. Of course, each firm, in deciding its output automatically induces price changes in the market. However, when both firms finally decide their outputs, price will be determined by the market. No changes in inventories are allowed in this model.

**The steps may be outlined as follows (K. J. Cohen and R. M. Cyert, Theory of the Firm):**

**1. Forecast of Competitors' Reactions:**

The forecast is basically a straightforward extrapolation of the past observed reactions of competitors.

**2. Forecast of Firm's Demand:**
This is based on an estimate of the demand function from past observations. Future demand is thus an extrapolation of the past sales of the firm.

**3. Estimation of Costs:**
The cost in the current period is assumed to be the same as in the past period. However, if the profit goal has been achieved over the past two periods, average unit costs are increased by a certain percentage to allow for slack payments.

**4. Specification of Goals of the Firm:**
These are aspiration levels. In this model profit is the only goal of the firm. The aspiration level of profits is some average of the profits of past periods.

**5. Evaluation of Results by Comparing Them to the Goals:**
From the information obtained in steps 1-3 we obtain a solution, i.e. an estimate of the level of output, price, cost and profits. These are compared to the target level of profits. If the goals are satisfied by this solution the firm adopts it. If the profit and other goals are not achieved the firm proceeds to step 6.

**6. If Goals are Not Attained the Firm Re-Examines the Estimate of its Costs:**
Re-examination starts with costs because this variable is under the direct control of the firm. It usually involves a cut in slack and other expenses.

**7. Evaluation of the New Solution by comparing it to Goals:**
If the new solution with the downward-adjusted costs leads to the target profits it is adopted. If not, the firm proceeds to step 8.

**8. If Goals are Not Attained the Firm Re-Examines the Estimate of its Demand:**
The re-examination consists in considering possible changes in the sales strategy (more market research, more advertising, more salesmen, etc). The result is an upward adjustment of the initial estimate of demand.

**9. Evaluation of the New Solution by comparing it to Goals:**
If the new solution with the revised costs and demand estimates attains the target profits, it is adopted. If not, the firm proceeds to step 10.

**10. If Goals are not met the Firm Readjusts Downwards its Aspiration Levels:**
If with the revision of costs (in step 6) and of demand (in step 9) the goals are not attainable, the firm readjusts downwards its aspiration levels. The firm has multiple goals (although only one explicitly appears in the above model), which take the form of aspiration levels the firm is a satisfice rather than a maximiser. The goals change over time depending on past

attainments, aspirations, demands of groups, and expectations. The criterion of choice for goal-setting is that the alternative selected meets the demands (goals) of the coalition.

The organisation seeks to avoid uncertainty. The market-originated uncertainty is avoided by undertaking information searches, by avoiding long-term planning, by following 'regular procedures and a policy of reacting to feedback information rather than of forecasting the environment. The competitor-originated uncertainty is avoided by creating a 'negotiated' environment, that is, by some sort of collusive behaviour.

**Comparison:**
The behavioural theory has contributed to the development of the theory of the firm in several respects. Its main contributions are: firstly, the insight into the process of goal-formation and the internal resource allocation, and secondly, the systematic analysis of the stabilizing role of 'slack' on the activity of the firm.

The behavioural theory deals with the allocation of resources within the firm, and the decision-making processes, an aspect neglected in the traditional theory. In the latter the firm was assumed to react to the all-powerful environment. The behaviourist school assumes that the firm has some discretion, and does not necessarily take the constraints of the environment as definite and impossible to change.

The traditional theory stressed the role of the market (price) mechanism for the allocation of resources between the various sectors of the economy, while the behavioural theory examines the mechanism of the resource allocation within the firm. Clearly the two theories are complementary rather than substitutes. Actually various theorists have attempted to incorporate the behavioural aspects of Cyert and March's theory into their own models.

Cyert and March's definition of 'slack' shows that this concept is equivalent to the 'economic rent' of factors of production of the traditional theory of the firm. The contribution of the behavioural school lies in the analysis of the stabilising role of 'slack' on the activity of the firm. Changes in slack payments in periods of booming and depressed business enable the firm to maintain its aspiration levels despite the changing environment.

It should be pointed out that Cyert and March deal only with one form of slack, the managerial slack. Slack payments accruing to other members of the firm-coalition and their short-run and long-run implications for the performance of the firm are not examined

**Criticisms of the Cyert and March Theory:**
The behavioural theory has, however, serious shortcomings. The Cyert and March theory of firm has been severely criticized on the following grounds:

1. The behavioral theory relates to a duopoly firm and fails as the theory of market structures. It does not explain the interdependence and interaction of firms, nor the way in which the interrelationship of firms leads to equilibrium of output and price at the industry level. Thus, the conditions for the attainment of a stable equilibrium in the industry are not determined.

2. The theory does not consider either the conditions of entry, effects on the behavior of existing firms of and the threat of potential entry by firms.

3. The behavioral theory explains the short-run behavior of firms and ignores their long-run behavior. It cannot explain the dynamic aspects of inventions and innovations which are related to the long-run.

4. The behavior theory is based on the simulations approach which is a predictive technique. It is simply the products of behavior of the firm but does not explain it.

5. The behavioural theories basically provide a simulation approach to the complexity of the mechanism of the modern multigoal, multiproduct corporation. Simulation, however, is a predictive technique. It does not explain the behaviour of the firm; it predicts the behaviour without providing an explanation of any particular action of the firm.

6. The behavioural theories do not deal with industry equilibrium. They do not explain the interdependence and interaction of firms, nor the way in which the interrelationship of firms leads to an equilibrium of output and price at the industry level. Thus, the conditions for the attainment of a stable equilibrium in the industry are not determined. No account is given of conditions of entry or of the effects on the behaviour of established firms of a threat by potential entrants.

7. The behavioural theory, although dealing realistically with the search activity of the firm (in the sense that search is considered as problem-oriented), cannot explain the dynamic aspects of invention and innovation, which are by their nature long-run activities with long-run implications.

8. The 'plasticity' (readjustment) of the aspiration levels downwards whenever the set targets are not attained deprives the theory of objective criteria for the evaluation of 'satisfactory' performance. To judge whether the performance of a firm is satisfactory one should have a 'constant measuring-rod', that is, a well-defined set of (long-run) goals. If goals are readjusted downward whenever their attainment has not been achieved, how are we to judge the performance of the firm? The 'measuring-rod' behaves like an elastic ruler that stretches and shrinks, depending on the attainment or not of the aspiration (goals) initially set.

9. No exact predictions can be derived from the postulates of the behavioural theory. The acceptance of satisficing behaviour renders practically the theory into a tautological structure: whatever the firms are observed to do can be rationalized on the lines of satisficing.

10. The behavioural theory implies a short-sighted behaviour of firms. Surely the uncertainty of the market cannot be avoided by short-term planning. Most decisions require a long-term view of the environment.

11. The behavioural theory resolves the chore problem of oligopolistic interdependence by accepting tacit collusion of the firms in the industry. This solution is unstable, especially when entry takes place, a situation brushed aside by the behavioural theorists. Cyert and March based their theory on four actual case studies and two experimental studies conducted with hypothetical firms.

**Conclusion:** Despite these criticisms, the behavioural theory of Cyert and March is an important contribution to the theory of the firm which brings into focus "multiple, changing and acceptable goals' in managerial decision-making.

## 7.5 FULL COST PRICING PRINCIPLE

**Introduction:**
For many years, Chamberlin 's and Joan Robinson's price theory of monopolistic competition had come to be generally accepted. According to This theory, the firms were able to act atomistically on the principle of profit maximisation without fear of rivals' reactions. They fixed prices so as to maximise their profits and this they did by equating marginal cost to marginal revenue (MC= MR). Empirical studies made by Oxford economists under the leadership of Professors Hall and Hitch (Price Theory and Business Behaviour) showed that the firms did not use the marginalist rule (MC = MR) and that oligopoly was the main market structure in the business world. According to Hall and Hitch, the firms did not act atomistically or irrespective of what their rival firms did. Rather they are continuously watching' the reactions of the rival firms. The traditional theory could not adequately explain the oligopolistic interdependence.

In such a situation, the firms do not attempt to maximise short-run profits by acting on the marginalized rule (MC = MR) but aim at maximising long-run profits by acting on the average-cost principle, i.e., the firms do not set their price and output at the intersection of MC and MR curves but they set them at a level which covers the average variable cost, (AVC) and average fixed cost (AFC) and normal profit margin in the business in question. AVC + AFC + Normal Profit. Firms do not seek abnormal profits for fear of to accept the prevailing price and has no option, therefore, the question of profit maximization does not arise. In the case (If monopolistic competition and absolute monopoly, the entrepreneurs are in a position to fix their price and maximize their profits. But in the case (If oligopoly, profit maximization cannot be considered a valid assumption. The oligopolist has both the desire and the power to achieve a secure position. In such a market situation, therefore, the desire

for security rather than the desire for maximum profit rules the entrepreneur's mind.

**Full Cost Pricing Principle:**

In 1939, Hall and Hitch published some results of research undertaken at Oxford University and aiming at the investigation of the decision processes of businessmen in relation to government measures. Their study covered 38 firms out of which 33 were manufacturing firms ,3 were retail trading firms and 2 were building firms. Out of 33 manufacturing firms, 15 produced consumer goods, 4 intermediate products, 7 capital goods, and 7 textiles. The sample was not random, but included firms which may well be expected to belong to 'efficiently managed enterprises.'

Hall and Hitch sought information from them about the elasticity and the position of their demand, and their attempts to equate their estimated marginal cost and marginal revenue. The answers revealed that the majority of them apparently made no efforts, even implicitly, to estimate elasticities of demand or marginal cost. They did not consider them to be of any relevance to the pricing process.

On the basis of the empirical study, Hall and Hitch concluded that the majority of entrepreneurs under oligopoly base their selling prices upon, what they call, 'full cost' and including an allowance of profit, and not in terms of the equality of marginal cost and marginal revenue at all.

Thus, a price based on full average cost is the 'right price', the one which 'ought to be charged', based on the idea of 'fairness to competition' under oligopoly. But what is full cost? Full cost is full average cost which includes average direct costs (AVC) plus average overhead costs (AFC) plus a normal margin for profit: Thus price, P = AVC + AFC + profit margin (usually 10%).

**According to Hall and Hitch, there are certain reasons which induce firms to follow the full-cost pricing policy:**

(i) Tacit or open collusion among producers;

(ii) Failure to know consumers' preferences;

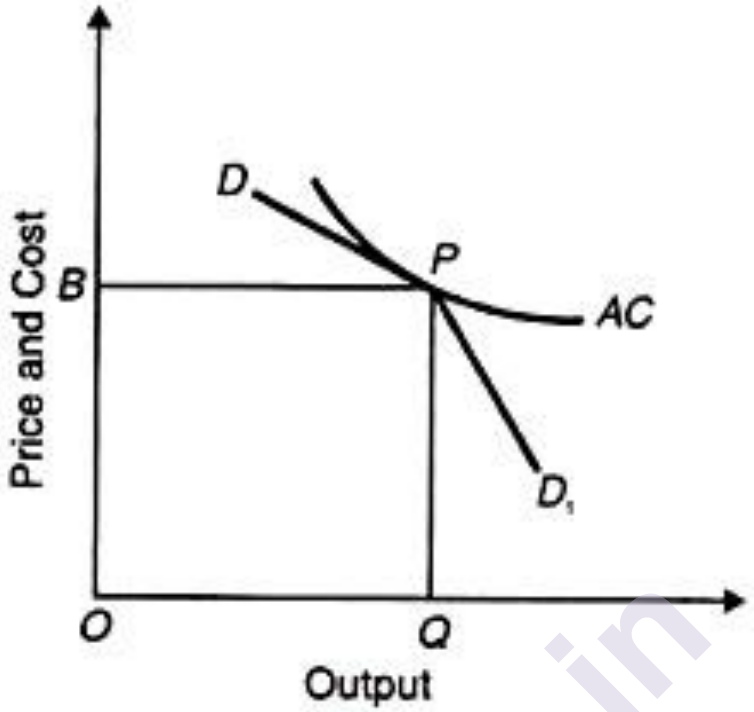
(iii) Reaction of competitors to a change in price;

(iv) Moral conviction of fairness; and

(v) Uncertainty of effects of price increases or decreases. All these reasons prevent oligopolistic producers from setting a price other than the full-cost price.

Thus, firms set their price on the basis of the full-cost principle and sell at that price whatever the market takes. They observed that prices were sticky in the oligopoly market despite changes in demand and costs. They explained the stickiness of prices in terms of the kinked demand curve. The kink occurs at the point where the price QP (= OB) fixed on the full-cost principle actually stands in Figure 7.15

**Figure No. 7.15**




Any increase in the price above it, will reduce the firm's sales, for its competitors will not follow it in raising their prices. This is because the PD portion of the kinked demand curve is elastic. On the other hand, if the firm reduces the price below QP, its competitors will also reduce their prices.

The firm will increase its sales but its profits will be less than before. This is because the $PD_1$ portion of the curve is less elastic. Thus, in both the price-raising and price-reducing situations, the firm will be a loser. It would, therefore, stick to the price QP so long as the prices of the direct factors of production (i.e., raw materials, etc.) remain unchanged.

As the AC curve falls over a large range of output, price varies inversely with output. The smaller the level of output, the higher will be the average cost and the higher the price of the product. But Hall and Hitch rule out the possibility of oligopoly firms producing small outputs and charging higher prices.

**They give three reasons for this;**

(a) Oligopoly firms prefer price rigidity,

(b) They cannot raise the price because of the kink, and

(c) They want to "keep the plant running as full as possible, giving rise to a general feeling in favour of price concessions".

**Hall and Hitch mention two exceptions to this phenomenon of a rigid price:**

(i) If the demand decreases much and remains so for some time, the price is likely to be reduced in the hope of maintaining output. This is likely to

happen when the lower portion of the demand curve becomes more elastic. The reason for this price- cut is when one firm becomes panicky and reduces its price; it forces others to cut their prices,

(ii) Any circumstances which lower or raise the AC curves of all firms by similar amounts due to changes in factor prices or technology are likely to lead to a revaluation of the full-cost price QP (= OB). But there is no tendency for prices to fall or rise more than the wage and raw material costs.

**The Andrews Version:**

The Hall-Hitch explanation is based on the presumption that the price to be charged in the oligopoly market is pre-set by the firm. Further, the kinky demand curve complicates the analysis. In order to simplify the exposition, we give a modified version of the full-cost pricing by Prof. Andrews.

Prof. Andrews in his study Manufacturing Business, 1949, explains how a manufacturing firm actually fixes the selling price of its product on the basis of the full cost or average cost. The firm finds out the average direct costs (AVC) by dividing the current total costs by current total output. These are the average variable costs which are assumed to be constant over a wide range of output.

In other words, the AVC curve is a straight line parallel to the output axis over a part of its length if the prices of direct cost factors are given. The price which a firm will normally quote for a particular product will equal the estimated average direct costs of production plus a costing-margin or mark-up.

The costing-margin will normally tend to cover the costs of the indirect factors of production (inputs) and provide a normal level of net profit, looking at the industry as a whole.

The usual formula for costing-margin (or mark-up) is,

$$M = P\text{-}AVC/AVC \ldots\ldots. (1)$$

Where M is mark-up, P is price and AVC is the average variable cost and the numerator P-AVC is the profit margin. If the cost of a book is Rs. 100 and its price is Rs. 125,

$$M = 125\text{-}100/100 = 0.25 \text{ or } 25\%$$

If we solve equation (1) for price, the result is

$$P = AVC\ (1 + M) \ldots\ldots. (2):$$

The firm should set the price

$$P = \text{Rs. } 100\ (1 + 0.25) = \text{Rs. } 125.$$

Once this price is chosen by the firm, the costing-margin will remain constant given its organisation, whatever the level of its output. But it will tend to change with any general permanent changes in the prices of the indirect factors of production.

Depending upon the firm's capacity and given the prices of the direct factors of production (i.e., wages and raw materials), price will tend to remain unchanged, whatever the level of output. At that price, the firm will have a more or less clearly defined market and will sell the amount which its customers demand from it.

**But how is the level of output determined?**
**It is determined in any of the three ways:**

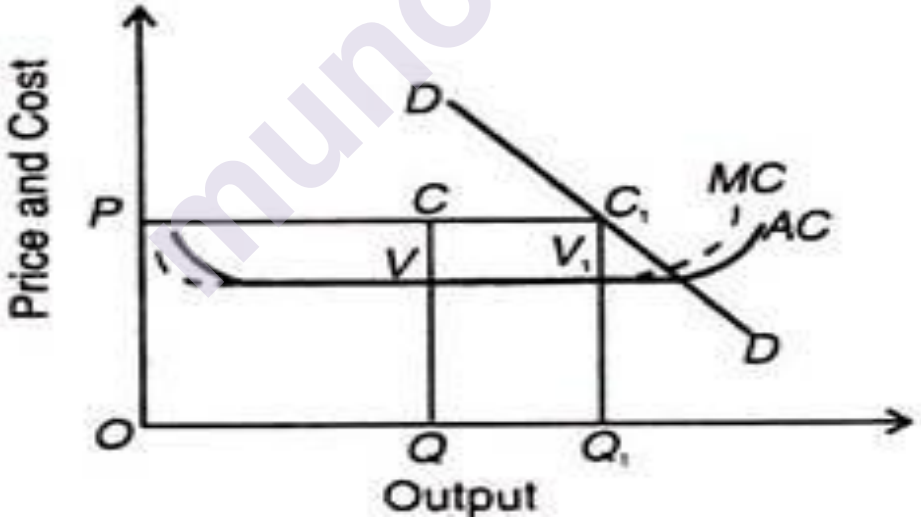
(a) As a percentage of capacity output; or

(b) As the output sold in the preceding production period; or

(c) As the minimum or average output that the firm expects to sell in the future.

If the firm is a new one, or if it is an existing firm introducing a new product, then only the first and third of these interpretations will be relevant. In these circumstances, indeed, it is likely that the first will coincide roughly with the third, for the capacity of the plant will depend on expected future sales.

**Figure No. 7.16**




The Andrews version of full-cost pricing is illustrated in Figure 4 where AC is the average variable or direct costs curve which is shown as a horizontal straight line over a wide range of output. MC is its corresponding marginal cost curve.

Suppose the firm chooses OQ level of output. At this level of output, QC is the full-cost of the firm made up of average direct costs QV plus the costing-margin VC. Its selling price OP will, therefore, equal QC.

The firm will continue to charge the same price OP but it might sell fig. 4 more depending upon the demand for its product, as represented by the curve DD. In this situation, it will sell $OQ_1$ output. This price will not be altered in response to changes in demand, but only in response to changes in the prices of the direct and indirect factors.

The following are advantages to using the full cost-plus pricing method:

1. **Simple:** It is quite easy to derive a product price using this method, since it is based on a simple formula. Given the use of a standard formula, it can be derived at almost any level of an organization. The concepts involved are familiar to businessmen and accountants.
2. **Price-Setting:** Average cost rule facilitates price-setting in multiproduct firms. In these firms acquisition of information on price elasticities for all productes is both difficult and costly.
3. **Likely profit:** As long as the budget assumptions used to derive the price turn out to be correct, a company is very likely going to earn a profit on sales if it uses this method to calculate prices.
4. **Justifiable:** In cases where the supplier must persuade its customers of the need for a price increase, the supplier can show that its prices are based on costs, and that those costs have increased.
5. **Classical Method:** It is the classical method of charging a price for a commodity. It is also a logical way of maximising long-run profit.
6. **Firms Ideal:** It is an ideal which the firms aim at. Covering the cost of production and earning a certain predetermined percentage of profit should be the objective even if it could not be fully achieved.
7. **Fair Price:** Price based on cost of production are considered fair for producers as well as consumers.
8. **Stop Frequent Changes:** Full cost piecing method can avoid frequent changes in price, Consumers do not appreciate changes in prices which occur frequently.
9. **Most preferred:** In reality market is uncertain and knowledge is incomplete making the market imperfect. Under these circumstances business people prefer a stable price based on full cost.

**Criticism:**

**The full-cost pricing theory has been severely criticised on the following grounds:**

**(1) Not free from profit maximisation:**
Critics like Robinson and Kahn have pointed out that the full-cost pricing theory is not free from the elements of profits maximisation which entered into the pricing decisions of many of the firms investigated by Hall and Hitch.

**(2) Whose full cost?**
One of the weaknesses of the theory is that it fails to point out the firm whose full cost will determine the price in the oligopoly market that will be followed by the other firms.

**(3) Firms follow independent price policy:**
The full-cost pricing theory is criticised for its adherence to a rigid price. Firms often lower the price to clear their stocks during a recession. They also raise the price when costs rise during a boom. Therefore, firms often follow an independent price policy rather than a rigid price policy.

**(4) Circular relationship:**
If fixed costs of a firm form a large proportion of its total cost, a circular relationship may arise in which the price would rise in a falling market and fall in an expanding market. This happens because average fixed cost per unit of output is low when output is large, and when it is small, average fixed cost per unit of output is low.

**(5) Profit margin a vague concept:**
Moreover, the term 'profit margin' or 'costing margin' is vague. The theory does not clarify how this costing margin is determined and charged in the full cost by a firm. The firm may charge more or less as the just profit margin depending on its cost and demand conditions.

As pointed out by Hawkins, **"The bulk of the evidence suggests that the size of the 'plus' margin varies: it grows in boom times and it varies with elasticity of demand and barriers to entry."**

**(6) Naive method:**
This pricing method is naive because it does not explicitly take into account the elasticity of demand. In fact, where the price elasticity of demand for a product is low, the cost plus price may be too low, and vice versa.

**(7) Not for perishable goods:**
This method cannot be used for price determination of perishable goods because it relates to the long period.

**(8) Full-cost pricing principle not strictly followed:**
Empirical studies in England and the U.S. on the pricing process of industries reveal that the exact methods followed by firms do not adhere strictly to the full- cost principle. The calculation of both of average cost and the margin is a much less mechanical process than is usually thought.
As a matter of fact, businessmen are reluctant to tell economists how they calculated prices and to discuss their relations with rival firms so as not to endanger their long-run profits or to avoid government intervention and maintain good public image.

**(9) Firms follow marginal principles;**
Prof. Earley's study of the 110 'excellently managed companies in the U.S. does not support the principle of full-cost pricing. Earley found a widespread distrust of full-cost principle among these firms. He reported that the firms followed marginal accounting and costing principles, and the majority of them followed pricing, marketing and new product policies.

**Conclusion:** Average-cost rules of pricing are useful for avoiding uncertainty and to 'co-ordinate' the market.

## 7.6 SUMMARY

In this way the module explains us the difference between the traditional theories and the new theories explaining the role of managers in the business activity. It explains us that how the managers try to expand the welfare of the labourers.

## 7.7 QUESTIONS

Q1. Write an explanatory note on marris model of managerial enterprise.

Q2. Explain Williamson's model of managerial discretion.

Q3. Write an explanatory note on Williamson's managerial Utility function.

Q4. Write a note on Behavioural Theory of firm.

Q5. Explain the behavioural theory of Cyert and March.

Q6. Explain in details the principle of full-cost pricing.

## 7.8 REFERENCES


- Gravelle H. and Rees R.(2004) : Microeconomics., 3rd Edition, Pearson Edition Ltd, New Delhi.
- Gibbons R. A Primer in Game Theory, Harvester-Wheatsheaf, 1992
- A. Koutsoyiannis : Modern Microeconomics
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.
- Varian H (2000): Intermediate Microeconomics: A Modern Approach, 8th Edition, W.W.Norton and Company
- Varian: Microeconomic Analysis, Third Edition
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.


*****

# 8

# ALTERNATIVE THEORIES OF THE FIRM-II

**Unit Structure**



## 8.0 OBJECTIVE

- To provide clear understanding of the concepts of Existence, Purpose and Boundaries of the firm and their importance.
- You will also learn the concepts of economies of scale
- To analyse the concept of resource-based theory
- Studying the importance of the knowledge in todays dynamic and competitive world.
- To study the concepts related to the Knowledge based Theory
- To understand the meaning of transaction cost

## 8.1 INTRODUCTION

Maximisation of profit is the main objective of the firm. But along with that the firm also tries to maximise several other objectives. Resources play a very important role in the growth of any firm. The Knowledge-based view of the firm is a recent extension of the Resource- based view of the firm very adequate to the present economic context. Knowledge is considered to be a very special strategic resource that does not depreciate in the way traditional economic productive factors do, and can generate increasing returns. The nature of most knowledge-based resources is mainly intangible and dynamic, allowing for idiosyncratic development through path dependency and causal ambiguity, which are the basis of the mechanism for economic rent creation in the Knowledge-based view of

the firm. Firms exist to simplify and reduce the transactional costs of coordinating economic activities

## 8.2 EXISTENCE, PURPOSE AND BOUNDARIES OF FIRMS

**What is a Firm?:**
A firm is a business organisation (which can take form as a corporation, partnership, Limited Liability Company (LLC) among others) which transforms inputs into outputs for profit. This can be in the form of goods (such as laptops or French fries), services (such as gardening or cleaning), or both (restaurants where we pay for the food we order, but also for the experience). Firms typically embody some kind of institutional structure with the management of the firm having both a set of objectives and a strategy with the goal of maximising profits.

Most firms, in the way we talk about in economics, are assumed that what a firm does maximises its profits. In general, that's true - it is in most legislation, and it is also required by law that a public traded company, which has shareholders, has an obligation to their shareholders who expect this firm would act in the best interest of the shareholders. There are also some other types of firms also such as

**Social enterprises:** Social enterprises are firms that exist for the purpose of maximising well-being and social impact, rather than profit. These firms may still be involved in buying and selling, as well as transforming inputs to outputs. In the business process, social enterprises try to achieve different purposes. For example, social enterprises may try to employ people from disadvantaged backgrounds who have difficulties in finding jobs otherwise, so they are paying back society through employment, or social enterprises may use their profits to help the society. In a sense, social enterprises are a bit like charities which don't get donations but instead use a firm as a vehicle to generate profits which are then donated to people in need. Given that a social enterprise is set up to maximise it's social impact, which it typically uses it's profits to drive, social enterprises are often still profit maximising firms.

**Government-owned companies:** In general, they serve the public and also have certain obligations like the Universal Service Obligation. Even if it is not profitable to serve certain persons because, for instance, they live very far and the outback, it might be required for those government-owned companies to serve those people, which they wouldn't necessarily do so if they were private companies due to the lack of profitability.

But in general, we deal with firms that maximise, or at least try to maximise their profits. As profits should be maximising at a long-term vision which can't be seen from now, expectations are needed to estimate the prospects of firms. Market prices of publicly traded firms are expected

to reflect those long-term visions, and people can deduce that if the share price of a company goes up, the firm is doing well in terms of its long-term profitability.

The theory of the firm consists of a number of economic theories that explain and predict the nature of the firm, company, or corporation, including its existence, behaviour, structure, and relationship to the market.

In simplified terms, the theory of the firm aims to answer few questions such as,

1. Existence. Why do firms emerge? Why are not all transactions in the economy mediated over the market?
2. Boundaries. Why is the boundary between firms and the market located exactly there with relation to size and output variety? Which transactions are performed internally and which are negotiated on the market?
3. Organization. Why are firms structured in such a specific way, for example as to hierarchy or decentralization? What is the interplay of formal and informal relationships?
4. Heterogeneity of firm actions/performances.What drives different actions and performances of firms?
5. Evidence. What tests are there for respective theories of the firm?

Firms exist as an alternative system to the market-price mechanism when it is more efficient to produce in a non-market environment. For example, in a labor market, it might be very difficult or costly for firms or organizations to engage in production when they have to hire and fire their workers depending on demand/supply conditions. It might also be costly for employees to shift companies every day looking for better alternatives. Similarly, it may be costly for companies to find new suppliers daily. Thus, firms engage in a long-term contract with their employees or a long-term contract with suppliers to minimize the cost or maximize the value of property rights.

Why do firms exist?
Firms exist to simplify and reduce the transactional costs of coordinating economic activities (Ronald Coase "The Nature of the Firm" 1937). By utilising the principles of economies of scale and scope, firms are able to reduce the transactional costs of operating within the market. Larger firms reduce costs by more efficiently satisfying 3 major factors required in economic activities:

**1. Search & Information:** Firms can minimise search costs regarding things like marketing and advertising (e.g. it's easier for a university than individual lecturers to find at their best price the lecture halls, students,

etc, as the university is doing these things for all lectures and all degrees throughout the university).

**2. Bargaining & Decision making**: Firms can use enterprise bargaining to set a price for everyone compared to freelancers negotiating at different prices with different people.

**3. Policing & enforcement**: Firms have strong policies in place to maintain quality.

An example of this is a model working freelance who has to do all of their own advertising, marketing and management of finances. If the model worked under a booking agent, their jobs would be set up, transport organised and payment collected by the firm. Therefore, it would take the pressure and stress of the model to perform their job better. To further explain, firms are needed to set a price and create a market. By creating a market a firm is able to consolidate demands of a certain good and produce it altogether to achieve economies of scale. Firms might even be able to use their resources to aid in the production of other in-demand products, which then becomes a form of economies of scope.

Industries formed by different firms competing in the same market may face disruption due to the rise of a new technology which helps eliminate transaction costs and consequently reduces the need for firms. Examples:

- **Person-to-person car sharing:** Where people's idle cars are temporarily made available to people who need transport. This results in a significantly lower demand for car rental agencies as any person can make their vehicles available through the application and can avoid the logistics required by a large firm (people might get rid of the firms that are normally organising these activities, and they might have individual trades with each other without the firms' getting percentage cuts).
- **Airbnb:** Airbnb is disrupting the hotel industry through the use of new technology such as applications to connect homeowners with travellers.
- **Video streaming:** online video streaming like YouTube and Netflix are disrupting Television company, which changes the way of people watching programs.
- **Coursera:** Coursera is disrupting universities as it provides massive free online learning courses which allow learners to be more flexible in their learning.

However, these new technologies can help further lower transaction cost which is a benefit for consumers. Over time, some conventional firms in the industry might be eliminated, while some might learn from these new technologies and further improve their industry standards.

**Horizontal Boundaries of Firms:**
Horizontal boundaries refer to the quantity (Economies of Scale) and variety (Economies of Scope) of products that a firm produces. Economies of Scale and Scope exist whenever large-scale operations provide a cost advantage over smaller ones, such that the average variable cost per unit reduces as the quantity of output of a single product, or a variety of products produced in a single plant increases. Production processes that are Capital Intensive generally are more likely to display Economies of Scale or Scope than Labour or Resource Intensive processes. Capital Intensive processes tend to have a higher fixed to variable cost ratio which benefits more from improving production techniques.

**Economies of Scale:**
Economies of scale are the cost advantages that firms obtain based on their scale of operation, with the cost per unit of output decreasing as the scale of production increases. However, some of the companies will not take advantage of economies of scale, preferring differentiation over cost leadership. When a market is producing at a level of economies of scale, allocative efficiency within the market is achieved. This means the market is producing at perfect competition, reducing costs and profit.

**Sources of Economies of Scale:**
Economies of Scale exists if there are:

**1. Indivisibilities in Production and the Spreading of Fixed Cost:**
One of the common sources of Economies of Scale is the spreading of fixed costs over an even larger volume of output produced. Indivisibilities refer to the minimum level at which any element of production requires to operate. Indivisibilities exist when the minimum level of production is significantly larger for new entrants to be economically viable. This occurs when there are high setup costs, long-run fixed costs, and volumetric returns to scale or a combination of all three. Larger firms can take advantage of indivisibilities by spreading costs over a greater volume of production as well as having better access to capital markets (assuming imperfect access to firms). Indivisibilities exist when it is possible to do things on a large scale that cannot be done on a small scale. Some inputs cannot be scaled down below a certain minimum size, even when the level of output is minimal. In general term, there is minimum expenditure a firm must incur in order to commence production (e.g. with a small backyard farm, a tractor is still needed to reduce labour intensity because it is not possible to purchase 0.01 of a tractor). Therefore, the first unit produced requires a significantly higher level of investment than the subsequent units, with the increase in subsequent units produced, it allows for costs to be spread out, allowing for Economies of Scale. If the indivisible input is not overly specialised, the firm can diversify its line of products at a lower cost as opposed to the total cost of individual specialised enterprises. Indivisibilities also promote economies of scope. For example, when airlines add new routes, they utilise conveyancing.

**2. Specialisation:**
Specialisation occurs when workers assigned to specific production tasks, increase in productivity and efficiency over time, allowing them to benefit from the lower average costs per increased in output. In order to conduct specialization, firms must be ready to make substantial investments, however, reluctance for firms will occur unless the present or forecasted demand justifies the volume to utilize specialization. From Adam Smith' theorem, it has stated that the division of labour is restricted to the span of the market (1) As the markets increase in size, economies of scale will enable the utilization of specialisation in productions, (2) The larger markets with volume advantages will support an arrangement of specialised activities

**3. Inventories:**
By carrying inventories, firms who conduct high volumes of business are able to maintain a lower ratio of inventory to sales. By buying in bulk, moving and storing big volumes of inventory reduces the overall cost per unit, hence Economies of Scale. Additionally, consolidation of inventories reduces costs associated with stock-outs and lost sales. There are various incentives for firms to possess inventories (1) Avoid stock-outs and lost sales (Safety stock is essential due to the uncertainty in the forecasts of sales. The added accuracy to the forecast, the fewer safety stock is required) (2)Avoid adversely influencing customer commitment, (3)Assuring no setbacks occur in the production process. However, by taking onto excess inventories, there will be consequences attached to such action as (1) Opportunity cost of cashflow restricted in inventory, (2) Rent, depreciation, insurance needed for inventory storage (3) Cost of deterioration and obsolescence of the inventories

**4. Large volumes of input purchases:**
Firms that purchase relatively greater quantities of inputs may obtain discounts from suppliers. Reasons for this include lower negotiating costs with a single supplier as opposed to multiple suppliers, suppliers benefiting from an association with reputable firms purchasing the inputs, and the security of confidentiality dealing with a single supplier can all induce discounting. Additionally, suppliers that rely on large purchases from a few firms are more inclined to discount, as they are risk-averse to losing the firms they supply.

**5. The Cube Square Rule:**
The mathematical concept which can be addressed to explain economies of scale. The Cube-Square Rule expresses the relationship between volume and the surface area. It implies that an increase in volume will incur an increase in surface area proportionally. This is another source of economies of scale. For many manufacturing processes, the capability of the machine to produce is related to the volume of the production vessel, and the total cost of production is closely associated with the surface area of the vessel. It is likely to have low average cost per unit by increasing

the capacity of production of the plant and decrease the ratio between the surface area and the volume of the production vessel.

**6.** Many processes are dealt with in volume but their costs are associated with an area (i.e. Storage). As the volume of a vessel increases by a given proportion, the surface area then increases by less than this proportion. In various production processes, production capacity is to be found proportional to the volume of the production vessel while the total cost of producing at capacity is proportional to the available surface area of the vessel. This concludes that as capacity increases, the average cost of producing at capacity will decrease because the ratio of surface area to volume decreases. For example, shipping of chairs in a shipping container. By stacking the chairs on top of each other, the capacity of chairs increases, hence decreasing the cost of shipping per surface area, achieving Economies of Scale.

**7. Marketing costs:**
Advertising has a certain fixed cost to all firms; therefore, larger firms are able to spread this cost over a relatively larger number of potential customers and can better adapt production to changes in demand from advertising campaigns in comparison to smaller firms. These fixed costs are similar for national firms as they are for regional firms. Firms with a greater scope of product offerings benefit from umbrella branding, which influences customer perception for all of the brand's products despite a campaign focusing on a single offering.

**8. Other sources:**
Other sources of Economies of Scale include labour specialisation, more efficient inventory management due to predictable customer demand and industries that encounter the cube square rule – where processes are volume related but costs are area related. A reduction in per-unit costs occurs in the short run when fixed costs are spread over increased production through better utilisation of a production plant's given capacity. In the long run this is represented by improvements in technology or increases in a plant's total production capacity, altering the dynamic of a firm's fixed to variable cost ratio.

**9. The network effect:**
The network effect is a unique source of Economies of Scale, which arises when customers experience greater benefit from using a product as a result of more people using it. For instance, Facebook provide the same value as a diary without the social function of interacting with other users. The utility and value of Facebook is higher than a diary is due to having increasingly high volume of users. The resulting 'demand-side' of Economies of Scale has a network effect if it benefits other adopters of the product (total effect) and incentivises others to adopt the product (marginal effect).

**Types of Economies of scale:**

**1. Internal:** Internal economies are factors and capabilities that are unique to and can be controlled by an organization that at minimal costs, can produce in large quantities. The big operational and financial size of an organization usually means they can take advantage of internal economies.

**2. External:** External economies result from advantageous conditions coming from outside the organization or, within an entire industry or economy. External economies mean that as an industry or sector grows, the average cost of doing business falls.

**Short-Run Economies of Scale:**
The reductions in unit costs are related to spreading fixed costs for a firm of a given size. Short-run economies of scale occur because of firms utilising a plant of a given capacity. For short-run economies of scale, it is assumed that there are fixed costs and the short-term average cost curve has a U-Shape [4]. The average cost in the short run is calculated by taking the total cost and dividing by output at each different level of output. Average cost shows that firms can earn profits given the market price.

**Long-Run Economies of Scale:**
The reductions in unit costs are caused by a firm switching from a low fixed/high variable cost plant to a high fixed/low variable cost plant. This happens when new technology is adopted by firms or when plant sizes are increased. For the long-run economies of scale, the average cost curve us more downward-sloping and it assumes that all factors/variables of production could change .

**Economies of Scope:**
Economies of scope happen when manufacturing one good causes the reduction of the production cost of another related product. As a result, which the marginal cost or the long-run average of a company decreases due to the production of complementary goods and services. Consequently, economies of scope are described by variety.

Economies of scope can be achieved when the cost of producing two different products together is less costly when a single firm produces them instead of tow separate firms. (ie. $C(q1,q2) < C(q1,0) + C(0,q2)$ where q1 the production level of good one and q2 is the production level of good two). This may occur when two products are complementary in their use to each other, when they have complementary production processes or when they share the same inputs to production.

**Learning economies:**
There is a learning economy where costs fall with experience. The learning economy can not directly expand the size of the company, but it can contribute to the success of the company. This stems from the fact that over time, managers and employees become more efficient in tasks. while

managers become better at allocating resources and scheduling production processes.

**Economies of Scale and Scope:**
Economies of both Scale and Scope present whenever large-scale production, distribution, or retail processes provide a cost advantage over smaller processes. In general, capital-intensive production processes are more inclined to demonstrate economies of scale and scope as compared to other labour or materials intensive processes. By allowing cost advantages, economies of scale and scope will not only influence the magnitude of firms and the structure of markets, but they will, too, configure critical business strategy arrangements, with an example of the possibility of the merging of independent firms and the likelihood of a firm achieving a long-term cost advantage. Economies of scale and scope are used to help cut a firm's operational costs. They occur when a firm experiences a cost advantage from implementing large-scale production over smaller processes. Economies of scope deal with average total cost of production of multiple goods while economies of scale are concerned with cost advantage that occurs due to the increased production of a single good. Economies of scope occur if it is possible for a firm to produce more than one good with the same resources, to increase the range of products they produce, while saving money on production costs, as opposed to producing the same amount of output with different resources. Economies of scale only occur with the indivisibilities, or the ability to manufacture products on a large scale that can't be manufactured on a smaller scale. Indivisibilities include returns to scale, long-run fixed costs and setup costs, costs that would be too expensive to maintain production if only a small volume of output was being produced.

**Differences between Economies of Scope and Economies of Scale:**
The economy of scope and economy of scale are two different concepts used to help cut a company's costs. Economies of scope focus on the average total cost of production of a variety of goods, whereas economies of scale focus on the cost advantage that arises when there is a higher level of production of one good. Economies of scale are reductions in average costs because production volume increases; whereas, economies of scope are reductions in average costs because the number of good produced increases.

**Differences:**

**1) Economies of scale**: firms reach a point of production where the cost of it no longer increases (bulk production). It is an old concept used in business and economics. This reduces the cost of one product. It consists in producing one type of product in bulk. The strategy behind is the standardization of the product. It uses a large number of resources because of bulk production.

**2) Economies of scope:** firms produce a variety of products and their cost of production gets reduced. It is a new term in business economics. This reduces the cost of multiple products. It consists of producing multiple products under the same operation. The strategy behind economies of scope is the diversification of products. It uses fewer resources because firms produce multiple products under one operation.

**Horizontal Mergers:**
Horizontal mergers have very high potential to have anticompetitive effects. This is because the total number of firms is reduced by one. Any potential increase in market power (ability of firm to raise prices above marginal cost) of a single firm must be balanced against any socially beneficial cost savings. Real world mergers can be very complex and require a number of steps to assess their viability.

**1. Market Definition:** this can be defined by the product, geography, product function, customers etc. Another way is the SSNIP test which refers to a 'small but significant non-transitory increase in price', this method helps define the market a firm operates in by assessing its market power.

**2. Safe Harbours:** mergers are significantly less likely to have negative effects on competition if post-merger market concentration is low. Market concentration can be determined by the Herfindahl index (HHI)

**3. Effect of Merger on Existing Competition:** evaluation of the competitive nature of the market, taking into account: the type of competition (price, quantity, fixed capacities), conduct of firms (coordinated?) and product differentiation

**4. Effect of Merger on Potential Competition:** possibility of entry deterrence/predation with or without merger, supplier relations and alternative technologies/networks

**5. Other Competition Factors:** changes in market powers of buyers and suppliers, scope for efficiency defence

**Vertical Boundaries of the Firm:**
Vertical boundaries of the firm refers to how much control the firm has over its industry operations, such as the production and distribution of their good or service.

Vertical integration can be divided into two streams – forward integration and backward integration.

In forward integration, companies will control their downstream counterparts, in order to increase control over the supply chain. For example, a gas mining company may own an energy power plant.

Backward integration refers to when companies seek to control their upstream counterparts, in order to increase control over the final product. For example, a chocolate manufacturer may seek to own cocoa farms.
Why not use the market for supplying inputs?

**Benefits of using the market:**

- Firms can achieve economies of scale that in-house departments producing for their own needs cannot - specialised firms will typically produce more than an in house department will
- Discipline of the market forces efficiencies on firms. That is, competition tends to promote effectiveness and quality. Relying on an in house department that meets the bare minimum requirements, will not have the same level of innovation & quality as an external firm.

**Costs of using the market:**

**1. Hold-up problem:** Is an issue of imperfect contracts - that is where negotiations/changes in circumstances can result in time delays or increased costs. This raises the costs of transacting market exchanges.

- It is argued that the possibility of hold-up can lead to underinvestment in relationship-specific investments and hence to inefficiency. For example, one supplier has an exclusive contract to supply body parts for the cars of General Motors. The supplier can hold up General Motors by increasing the price for the additional parts produced if exceeding demands occur.
- It can lead to difficult contract negotiations and more frequent renegotiations
- It can lead to distrust between corporations

**2. Difficulties in coordination:** External firms are harder to control than internal departments. This in turn can raise costs with bottlenecks in the production flow. The failure of one firm to deliver supplies on time can lead to another factory being shut down.

**3. Security of private information:** Private information may be leaked when using the market. Leakages can result in firms' competitive advantage being compromised. An example of this is a patent or special know-how.

**4. Transaction costs in contracting:** Cost incurred during the process of purchasing and selling goods or services.

**Vertical Separation:**
Some firms may decide to develop looser relationships then complete full vertical integration. That is, instead of fully moving all production in-house, they will utilise a balance between internal departments and the market.

**Advantage of a looser relationship over full vertical integration**

1. Preservation of firm's independence
2. Avoidance of costs that may be associated with full vertical integration

**Examples of looser relationships:**

- Franchising: involves a specific contractual relationship/arrangement between franchiser & franchisee E.g., McDonald's, Hungry Jacks, 7-Eleven
- Networks of independent firms that are linked vertically & establish nonexclusive contracts or relationships with one another E.g. Grocery retailers and Metcash (grocery wholesaler)

**Other alternatives to vertical integration:**

**Tapered integration:** A mix of vertical integration and market exchange, making some inputs and buying the remaining portion from independent firms. Example: BMW uses some external market research along with in house market research. The advantage of Tapered Integration: Producing part of the production requested materials and input the rest of the materials from other companies in exists in the market. This will reduce the initial cost of capital and reduce the cost of misunderstanding the market price. Manufacturing some of the demand whist purchasing the rest from the market will not only increase the bargaining power of the company itself, and also threat the external suppliers to discipline the supply process and quality of the supplies. Disadvantage of Tapered Integration: The company may not achieve the economies of scale, because of the sufficiency of production will need both internal production and external suppliers to coordinate. Other than the loss of economies of scale, the tapered integration may incur higher coordination costs and freight in and out costs due to purchasing supplies from external suppliers. The efficiency of production process will also be a problem if coordination of supplies and internal production process does not collaborate.

**Joint Venture:** Where two or more parties decide to work together by pooling their resources with the goal of achieving a specific task or completing a certain business activity. However, the venture is separate from the other business interests and the two companies operate as one in the venture. Example: the creation of google earth was as a result of a joint venture between Google and NASA.

**Strategic Alliance:** A strategic alliance is where two or more firms work together to increase each other's performance. They operate in the same way as a joint venture however what makes them different is they operate as separate companies and don't require a legal contract. Example: ApplePay and Mastercard; Mastercard was the first to offer ApplePay this alliance means they benefit from sharing their users.

**Long term collaborative relationships:** At least two parties who agree to share resources, such as finance, knowledge and people to accomplish a mutual goal. Example: business relationships.

**Implicit contracts between firms:** Are a non-binding agreement voluntarily entered into in regard to future exchanges of goods and services. Example: an employer continues to offer employment given the employee remains sincere in not looking for another job and continues their duties

**Recently in Western countries:** This strategy forester the vertical disintegration and concentrate on create core competencies for companies, aiming to outperform other companies in within the market.

## 8.3 THE RESOURCE-BASED THEORY OR VIEW (RBT/RBV)

The resource-based view / theory (RBV) is a managerial framework used to determine the strategic resources a firm can exploit to achieve sustainable competitive advantage.

Barney's 1994 article "Firm Resources and Sustained Competitive Advantage" is widely cited as a pivotal work in the emergence of the resource-based view. However, some scholars argue that there was evidence for a fragmentary resource-based theory from the 1930s. RBV proposes that firms are heterogeneous because they possess heterogeneous resources, meaning firms can have different strategies because they have different resource mixes.

The RBV focuses managerial attention on the firm's internal resources in an effort to identify those assets, capabilities and competencies with the potential to deliver superior competitive advantages.

**Origins and background:**
During the 1990s, the resource-based view (also known as the resource-advantage theory) of the firm became the dominant paradigm in strategic planning. RBV can be seen as a reaction against the positioning school and its somewhat prescriptive approach which focused managerial attention on external considerations, notably industry structure. The so-called positioning school had dominated the discipline throughout the 1980s. In contrast, the resource-based view argued that sustainable competitive advantage derives from developing superior capabilities and resources. Jay Barney's 1991 article, "Firm Resources and Sustained Competitive Advantage," is seen as pivotal in the emergence of the resource-based view. A number of scholars point out that a fragmentary resource-based perspective was evident from the 1930s. Scholars suggest that the resource-based view represents a new paradigm, albeit with roots in "Ricardian and Penrosian economic theories according to which firms can

earn sustainable supranormal returns if, and only if, they have superior resources and those resources are protected by some form of isolating mechanism precluding their diffusion throughout the industry."

The RBV is an interdisciplinary approach that represents a substantial shift in thinking. The resource-based view is interdisciplinary in that it was developed within the disciplines of economics, ethics, law, management, marketing, supply chain management and general business.

RBV focuses attention on an organization's internal resources as a means of organising processes and obtaining a competitive advantage. Barney stated that for resources to hold potential as sources of sustainable competitive advantage, they should be valuable, rare, imperfectly imitable and not substitutable (now generally known as VRIN criteria). The resource-based view suggests that organisations must develop unique, firm-specific core competencies that will allow them to outperform competitors by doing things differently.

The resource-based view (RBV) of the organisation is a strategy for achieving competitive advantage that emerged during the 1980s and 1990s, following the works of academics and businessmen. Key theorists who have contributed to the development of a coherent body of literature include Birger Wernerfelt, Spender, Grant.Jay B. Barney, George S. Day, Gary Hamel, Shelby D. Hunt, G. Hooley and C.K. Prahalad.

The core idea of the theory is that instead of looking at the competitive business environment to get a niche in the market or an edge over competition and threats, the organisation should instead look within at the resources and potential it already has available.

According to RBV, it is significantly easier to exploit new opportunities using resources and competencies that are already available, rather than having to acquire new skills, traits or functions for each different opportunity. These resources are the main focus of the RBV model, with its supporters arguing that these should be prioritised within organisational strategy development.

Although the literature presents many different ideas around the concept of the resource-advantage perspective, at its heart, the common theme is that the firm's resources are financial, legal, human, organisational, informational and relational; resources are heterogeneous and imperfectly mobile and that management's key task is to understand and organise resources for sustainable competitive advantage.

**Concept:**
Achieving a sustainable competitive advantage lies at the heart of much of the literature in strategic management and strategic marketing. The resource-based view offers strategists a means of evaluating potential factors that can be deployed to confer a competitive edge. A key insight

arising from the resource-based view is that not all resources are of equal importance, nor do they possess the potential to become a source of sustainable competitive advantage. The sustainability of any competitive advantage depends on the extent to which resources can be imitated or substituted. Barney and others point out that understanding the causal relationship between the sources of advantage and successful strategies can be very difficult in practice. Thus, a great deal of managerial effort must be invested in identifying, understanding and classifying core competencies. In addition, management must invest in organisational learning to develop, nurture and maintain key resources and competencies. Resource- based theory contends that the possession of strategic resources provides an organisation with a golden opportunity to develop competitive advantage over its rivals.

In the resource-based view, strategists select the strategy or competitive position that best exploits the internal resources and capabilities relative to external opportunities. Given that strategic resources represent a complex network of inter-related assets and capabilities, organisations can adopt many possible competitive positions. Although scholars debate the precise categories of competitive positions that are used, there is general agreement, within the literature, that the resource-based view is much more flexible than Porter's prescriptive approach to strategy formulation. Identification, evaluation, development, protection, expansion etc. of resources becomes very much essential in strategic management process.

The key managerial tasks are:

1. Identify the firm's potential key resources.
2. Evaluate whether these resources fulfill the following criteria (also known as VRIN criteria:

- Valuable - they enable a firm to implement strategies that improve its efficiency and effectiveness.
    - Rare - not available to other competitors.
    - Imperfectly imitable - not easily implemented by others.
    - Non-substitutable - not able to be replaced by some other non-rare resource.

3. Develop, nurture and protect resources that pass these evaluations.

Given the centrality of resources in terms of conferring competitive advantage, the management and marketing literature carefully defines and classifies resources and capabilities. It is defined and explained as follows.

**Resources:** Barney defines firm resources as: "all assets, capabilities, organizational processes, firm attributes, information, knowledge, etc. controlled by a firm that enable the firm to conceive of and implement strategies that improve its efficiency and effectiveness."

**Capabilities**: Capabilities are "a special type of resource, specifically an organizationally embedded non-transferable firm-specific resource whose purpose is to improve the productivity of the other resources possessed by the firm."

**Competitive advantage:** Barney defined a competitive advantage as "when [a firm] is able to implement a value creating strategy not simultaneously being implemented by any current or potential competitors."

**Classification of Resources and Capabilities:**
Within an RBV model, there are two main types of resource (assets), which will likely be familiar to accountants and financial specialists. Firm-based resources may be divided into two main categories viz tangible or intangible.

**1. Tangible resources**: These are physical assets such as financial resources and human resources including real estate, property, raw materials, machinery, plant, inventory, brands, land, products and capital, patents and trademarks and cash. These are resources which can generally be bought easily on the market and thus offer little competitive advantage, as other organisations can also acquire identical assets quickly if they should like.

**2. Intangible resources:** This refers to items and concepts that have no physical value but can still claim to be owned by the organisation. These may be embedded in organisational routines or practices such as an organization's reputation, culture, knowledge or know-how, accumulated experience, relationships with customers, suppliers or other key stakeholders, trademarks or intellectual property which the organisation may possess. Some of these - e.g. reputation - are built up over a significant period of time, and is something which other competitors or comparable organisations cannot buy on the market. These will likely stay within the organisation and are their main source of competitive advantage. They are particularly valuable in resource-based view because they give companies advantages in using resources. For example, patents make it impossible for other firms to use their resources in the same way and brand might be the only thing differentiating the product from the competitor's.

**The resources are divided into two critical assumptions:**
**1. Heterogeneous:** This first major assumption is that resources, skills and capabilities must vary significantly from one organisation to another. It is the assumption that each company has different skills, capabilities, structure, resources and that makes each company different. Due to the different forms of employment and number of resources, organizations can design different strategies that promote competitiveness in the market. If these organisations had the exact same set of resources and individuals, they would not be able to employ varying strategies in order to compete

with one another, as other organisations would be able to follow them step-by-step (known as "perfect competition").

Perfect competition does not exist in the real world - companies may be exposed to the exact same competitive and external forces, but they are still able to formulate different strategies to compete with one another. Thus, RBV assumes that this is due to the varying values of their resources and skills.

**2. Immobile:** The second assumption of RBV is that resources are immobile, and thus unable to move freely from organisation to organisation (e.g., employee movement), at least over the short-term. Due to this, organisations are unable to quickly replicate the resources of rival organisations and therefore implement the same strategies. Intangible assets - knowledge, processes, intellectual property, etc. - are more likely to be 100% immobile than are tangible assets.

It is the assumption that is based on the resources that an organization owns are not mobile, in other words, at least in short terms, cannot be transferred from one company to another. Companies can hardly obtain the immobile resources of their competitors since those resources have an important value for companies.

- A resource is valuable up to which it helps a firm create unique strategies that capitalize on opportunities and diminishes threats. A resource is non-substitutable when alternative ways to gain the benefits the resource provides is impossible to get. A rare resource provides strategic advantages to the company which owns it.
- Competitors find it hard to duplicate resources that are difficult to imitate. Some of these are protected by various legal means, including trademarks, patents, and copyrights.
- Resource-based theory also focuses on the merit of an old saying “the whole is greater than the sum of its parts”. Strategic resources can be created by various strategies and resources, bundling them together in a way that cannot be copied. Distinguishing strategic resources from other resources is important. Cash is an important resource. Tangible goods, including car and home are also vital resources.

**From Resources to Capabilities:**

**Resources and capabilities may also be intraorganizational or interorganizational:**
While RBV scholars have traditionally focused on intraorganizational resources and capabilities, recent research points to the importance of interorganizational routines. Routines between organizations and the ability to manage interorganizational relationships can improve performance. Such collaboration capabilities are, in particular, supported by contract design capabilities. An efficient use of contracts in the management of interorganizational relationships can facilitate the transfer

of information, enhance organizational learning, and help develop relational capital.

- The tangibility of a firm's resource is an important consideration within resource-based theory. Tangible resources are resources that can have a physical presence. A firm's property, plant, and equipment, as well as cash, are tangible resources.
- In contrast, intangible resources are not physically present. The knowledge and skills of employees, a firm's reputation, and a firm's culture are intangible resources.

**Capabilities:** Capabilities are another key concept. Resources refer to what an organization owns, capabilities refer to what the organization can do. Capabilities often arise over time while the firm takes actions that build on its strategic resources. Some firms develop a dynamic capability, where a company has a unique ability of creating new capabilities to keep pace with changes in its environment.

Dynamic Capabilities of GE and Coca Cola: General Electric, for example, buys and sells firms to maintain its market leadership over time, while Coca-Cola is known for building new brands and products as the soft-drink market changes. Both of these firms are among the top fifteen among the "World's Most Admired Companies".

**The Importance of Marketing Mix:**

- Leveraging resources and capabilities to create desirable products and services is important. The marketing mix—also known as the four Ps of marketing—provides important insights into how to make customers convinced to purchase the goods and services.
- The real purpose of the marketing mix is not to cheat but actually to provide a strong combination among the four Ps (product, price, place, and promotion) to offer the customers a useful and persuasive message.

**VRIO Framework:**
Although possession of heterogeneous and immobile resources is crucial to organisational success, it is not alone if they wish to sustain this competitive advantage.

Barney (1991) identified a framework for examining the key properties of resources and organisations (VRIO). These criteria were altered later by other leadership thinkers, and the new acronym VRIO was developed. This stands for:

- **Valuable:** Resources are valuable if they can help to increase the value of the service or product supplied to customers or others reliant on the organisation. This can be improved by increasing differentiation, decreasing the cost of production, or other general modifications to improve the quality and worth of the service. Any

resources that do not meet this condition may lead to a competitive disadvantage.

- **Rare. Any resources :** both tangible or intangible - which can only be acquired by one or very few organisations, may be considered rare. If organisations have the same resources or capabilities, this can result in competitive parity.
- **Low Imitability:** If an organisation holds resources which are valuable or rare, they can at least achieve a competitive advantage in the short-term. However, to sustain this advantage the resources need to be costly to imitate or substitute, or else rivals may begin to close the gap by obtaining the same or similar resources.
- **Organised to capture value:** Resources do not necessarily convey a competitive advantage - if the organisation, its systems and its processes are not designed to exploit the resource to its fullest, then it cannot hope to gain a competitive advantage. This could refer to not utilising talented or knowledgeable individuals in the correct department or role, or not fully building campaigns that utilise the organisation's positive reputation, amongst many other examples.

Only when all of these factors are fulfilled can one gain a sustained competitive advantage, and can innovate and get ahead in the market. The process for maximising an advantage using the RBV should follow as such:

1. Identify the organisation's potential key resources
2. Evaluate whether the resources fulfil the VRIO criteria (using the flowchart below)
3. Develop and nurture the resources that pass these criteria

If organisational leaders do as such, the organisation should hypothetically be expected to pull ahead of rivals and to advance through new ground in the market.

**RBV and strategy formulation:**

Firms in possession of a resource, or mix of resources that are rare among competitors, are said to have a comparative advantage. This comparative advantage enables firms to produce marketing offerings that are either (a) perceived as having superior value or (b) can be produced at lower costs. Therefore, a comparative advantage in resources can lead to a competitive advantage in market position.

In the resource-based view, strategists select the strategy or competitive position that best exploits the internal resources and capabilities relative to external opportunities. Given that strategic resources represent a complex network of inter-related assets and capabilities, organisations can adopt many possible competitive positions. Although scholars debate the precise categories of competitive positions that are used, there is general agreement, within the literature, that the resource-based view is much

more flexible than Porter's prescriptive approach to strategy formulation. Though the original formulators of the RBV play down the importance of external activity within the market, Hooley et al. (1998) have suggested that the marketing paradigm and the RBV are not unreconcilable, and that external strategic planning is still important for success.

In an RBV-centric organisation, leaders should select strategies that best exploit internal resources relative to external opportunities and competition. This can involve many different strategic positions, due to the variety of forms which resources can take.

Hooley et al. suggest that there are six different competitive positions one can take when utilising a resource-based view of the organisation:

- Price positioning
- Quality positioning
- Innovation positioning
- Service positioning
- Benefit positioning
- Tailored positioning (one-to-one marketing)

These various strategies have been posited as being significantly less rigid than Porter's well-known competitive strategies, and depend entirely upon the resources available to the firm.

**Criticisms:**
A number of criticisms of RBV have been widely cited and are as follows:

- The RBV is tautological
- Different resource configurations can generate the same value for firms and thus would not be competitive advantage
- The role of product markets is underdeveloped in the argument
- The theory has limited prescriptive implications.

**Other criticisms include:**

- The failure to consider factors surrounding resources; that is, an assumption that they simply exist, rather than a critical investigation of how key capabilities are acquired or developed.
- It is perhaps difficult (if not impossible) to find a resource which satisfies all of Barney's VRIN criteria.
- An assumption that a firm can be profitable in a highly competitive market as long as it can exploit advantageous resources does not always hold true. It ignores external factors concerning the industry as a whole; Porter’s Industry Structure Analysis ought also be considered.

- Supporters of RBV posit that competitive advantage is best achieved by utilising present internal resources. However, this has drawn many critics within leadership and management, and other theories and frameworks such as the industrial organisation view (I/O), place more emphasis on strategic planning, regulatory policy and the activity of market competition.

- In reality, the likelihood is that significant amounts of an organisation's performance can be explained by both factors, though some studies have indicated that internal resources are indeed more important with regards to competitive advantage and performance overall.

- There are other critiques, however. The authors of RBV frameworks tell managers that they should find and develop high potential resources, using the VRIO framework; however, they do not suggest how this should be done, and in reality, there is often nothing that managers appear able to do to improve the resources available. What it does neglect to mention, is that leaders and managers have the capability to improve the processes and systems that create higher-value resources - which could over the longer-term have a more significant impact on the performance of the organisation.

- In addition - when in unpredictable markets such as the technology industry, innovations and new inventions can almost-instantly have a drastic effect on the value of resources. This can render previous activities to try and generate a sustainable advantage totally null - thus, RBV can be considered to only be a practical view when situated in a stable competitive environment. Some (e.g. Eisenhardt and Martin, 2000) have indicated that levels of organisational learning and adaptiveness are more crucial to success over the long term, though RBV can be an important model in the short term.

- Further critiques include the extreme rarity of resources that match the VRIO criteria, the limit of the VRIO criteria itself in determining value, the unclear and indeterminate nature of VRIO itself, and the ambiguous nature of the term "resources". The general concluding thought is that RBV can be useful for developing competitive advantage, particularly in the short-term, but should be considered in partnership with other frameworks and theories when performing long-term strategic planning.

## 8.4 KNOWLEDGE BASED THEORY OF FIRM

**Introduction:**

The knowledge-based theory of the firm considers knowledge as the most strategically significant resource of a firm. Its proponents argue that because knowledge-based resources are usually difficult to imitate and socially complex, heterogeneous knowledge bases and capabilities among

firms are the major determinants of sustained competitive advantage and superior corporate performance.

This knowledge is embedded and carried through multiple entities including organizational culture and identity, policies, routines, documents, systems, and employees. Originating from the strategic management literature, this perspective builds upon and extends the resource-based view of the firm (RBV) initially promoted by Penrose (1959) and later expanded by others (Wernerfelt 1984, Barney 1991, Conner 1991).

Although the resource-based view of the firm recognizes the important role of knowledge in firms that achieve a competitive advantage, proponents of the knowledge-based view argue that the resource-based perspective does not go far enough. Specifically, the RBV treats knowledge as a generic resource, rather than having special characteristics. It therefore does not distinguish between different types of knowledge-based capabilities. Information technologies can play an important role in the knowledge-based view of the firm in that information systems can be used to synthesize, enhance, and expedite large-scale intra- and inter-firm knowledge management.

**The knowledge-based theory:**

In the last two decades of the 20th century a resource-based theory of the firm has received attention as an alternative to the traditional product-based or competitive advantage. The resource-based perspective promises to improve understanding of strategy formulation also in firms, which are dependent on intangible resources, such as, the rapidly growing knowledge-based services and knowledge-intensive industries. Organizational knowledge presents a tremendous wealth creating potential.

Contrary to traditional and finite production factors, knowledge can generate increasing returns, through its systematic use. Knowledge presents very special characteristics that differentiate it from physical resources and contribute to the creation and sustainability of competitive advantage. Knowledge can be used simultaneously in several applications and still it does not devaluate. Organizational knowledge is such a marvellous substance, contrary to other resources, its utilization, under different forms, increases it, instead of decreasing it. Knowledge-based capabilities are considered to be the most strategically important.

ones to create and sustain competitive advantage A distinction was made between three epistemologies that guided the practice and research under an epistemological perspective: the cognitivist, the connectionist and the autopoietic. The cognitivist perspective assumes organisations to be open systems, which develop knowledge by formulating increasingly accurate "representations" of the world. The more data and information

organisations can gather the closer the representation will be. Hence most cognitivist perspectives equate knowledge with information and data.

According to the connectionist epistemology the organisation still "represents" its outside world, but the process of representation of reality is different. As in cognitivist epistemology information processing is the basic activity of the system.

Autopoietic epistemology provides a fundamentally different understanding of the input to a system. Input is regarded as data only. Knowledge is private concept related to "personal" knowledge. Autopoietic systems are both closed and open. Open to data, but closed to information and knowledge, both of which have to be interpreted inside the system. Autopoietic systems are self-referring, it is constructed within the system and it is therefore not possible to "represent" reality.

Knowledge defined as a "capacity-to act" is dynamic, personal and distinctly different from data (discrete, unstructured symbols) and information (a medium for explicit communication).

**A Knowledge-Based Theory for Strategy Formulation:**
The word "Strategy" is usually associated with activities and decisions concerning the long-term interaction of an organisation with its environment. While competitive-based and product-based strategy formulation generally makes markets and customers the starting point for the study the resource-based approach tends to place more emphasis on the organisation's capabilities or core competences.

A knowledge-based strategy formulation starts with the primary intangible resource: the competence of people. People are seen as the only true agents in business; all tangible physical products and assets as well as the intangible relations are results of human action, and depend ultimately on people for their continued existence. People are seen to be constantly extending themselves into their world by both tangible means, such as craft, houses, gardens and cars and intangible corporate associations, ideas, and relationships. These intangible extensions are called 'media'.

People can use their competence to create value in two directions: by transferring and converting knowledge externally or internally to the organisation they belong to. When the managers of a manufacturer direct the efforts of their employees internally, they create tangible goods and intangible structures such as better processes and new designs for products. When they direct their attention outwards, they will in addition to delivery of goods and money also create intangible structures, such as customer relationships, brand awareness, reputation and new experiences for the customers.

**Three 'Families' of Intangible Resources:**
The External structure can be seen as a family3 of intangible relationships with customers and suppliers, which form the basis for the reputation (image) of the firm. Some of these relationships can be converted into legal property such as trademarks and brand names. The value of such

intangible resources is primarily influenced by how well the company solves its customers´ problems, which involves an element of uncertainty.

Internal Structure can be seen or created when people direct their actions internally. The family of Internal Structure can be seen to hold patents, concepts, models, templates, computer systems and other administrative more or less explicit processes. These are created by the employees and are generally "owned" by the organisation. "culture" or the "spirit" can also be regarded as belonging to the internal structure.

The Individual Competence family consists of the competence of the professional/technical staff, the experts, the R&D people, the factory workers, sales and marketing – in short all those that have a direct contact with customers and whose work are directly influencing the customers view of the organisation.

knowledge transfers are different from tangible goods transfers. In contrast to tangible goods, which tend to depreciate in value when they are used, knowledge grows when used and depreciates when not used. Building up competence in a language or a sport requires huge investments in training and managerial competence takes a long time on-the-job to learn. If one stops speaking the language it gradually dissipates.

The manufacturing and transportation of physical goods from suppliers, via a factory to a buyer gave us the concept of the Value Chain. If we see the organisation as creating value from transfers and conversions of knowledge together with its customers the Value Chain collapses and the relationship should better be seen as a Value Network; an interaction between people in different roles and relationships who create both intangible value (knowledge, ideas, feedback, etc) and tangible value.

Individual Competence External Structure Internal Structure $ Knowledge transfers, knowledge conversions

**Figure No. 8.1**

**The Firm from a Knowledge-based Perspective4**

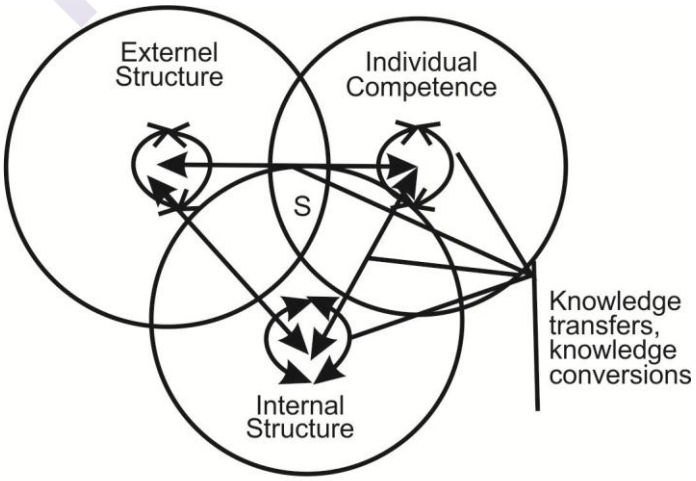



Figure 1 The Firm from a knowledge- based perspective

In contrast to the Value Chain the intangible value in a Value Network grows each time a transfer takes place because knowledge does not physically leave the creator as a consequence of a transfer. The knowledge I learn from you adds to my knowledge, but it does not leave you. Thus, from an organisational viewpoint the knowledge has effectively doubled. Knowledge shared is knowledge doubled. From an individual's point-of-view the perspective however, is different. Here knowledge shared may be an opportunity lost if the effect of the sharing becomes lost career opportunities, extra work and no recognition. Knowledge shared can be competitiveness lost. Fear of dismissal or competition are commonly cited reasons why individuals do not share what they know or what they create.

While the above primarily is concerned with transfer of existing (often hidden and/or underutilised knowledge), another issue is the creation of entirely new knowledge. Some have argued that new knowledge is created in the conversion of explicit/tacit knowledge from one type to another.

The strategy formulation issues are concerned with how to utilise the leverage and how to avoid the blockages that prevent sharing and creation of new knowledge. The key to value creation lies with the effectiveness of such transfers and conversions. The choice of the words "transfer" and "conversion" may suggest one-directional movements of knowledge. This is not the intention. Knowledge transfer between two individuals is a bidirectional process, which tends to improve competence of both and teamwork tends to be a cocreation of knowledge involving the whole team. Moreover, transfer of competence depends on conversion from tacit to explicit and back to tacit again in an endless spiral.

One feature of a knowledge-based theory of the firm is that it challenges perceptions about the boundaries of an organisation. What is indeed "the organisation" if customers and suppliers are included as families of the firm as in Figure 1? When the importance is placed on how effective the value creation is in the whole system, the issue of whether an individual is a formal employee, a customer, a contractor, a supplier or a customer becomes less of an issue as long as the relationship generates value. An ex-employee can for instance be more valuable as a customer than as an employee, a fact long exploited by the professional services firms.

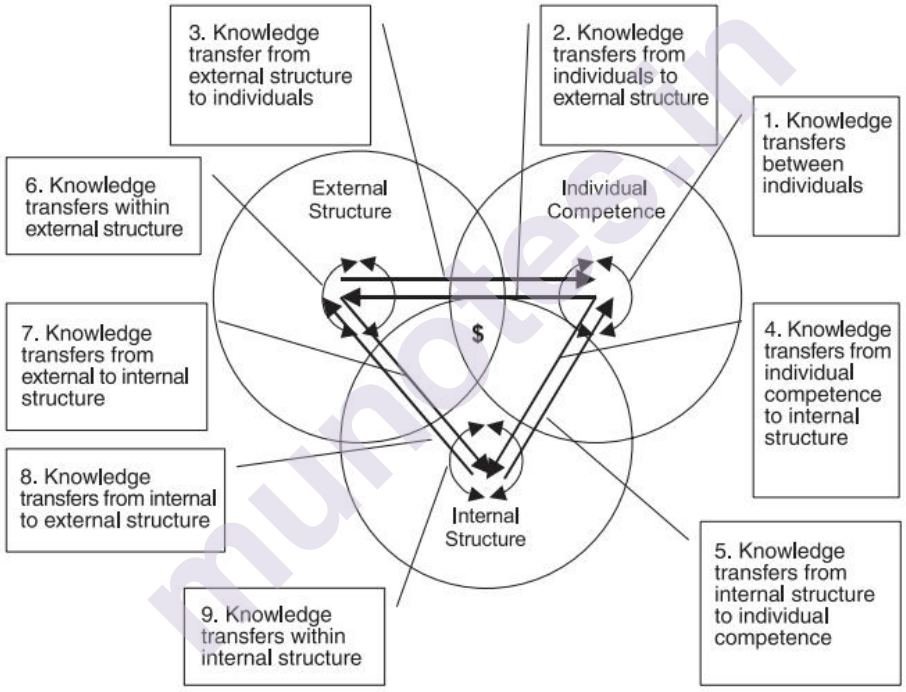
The Ten Knowledge Strategy Issues

From the framework above we can distinguish nine basic knowledge transfers/conversions, which have the potential to create value for an organisation. Activities that form the backbone of a knowledge strategy, are to be aimed at improving the capacity-to-act of people both inside and outside the organisation.

1. Knowledge transfers/conversions between individuals
2. Knowledge transfers/conversions from individuals to external structure

3. Knowledge transfers/conversions from external structure to individuals
4. Knowledge transfers/conversions from individual competence into internal structure
5. Knowledge transfers/conversions from internal structure to individual competence
6. Knowledge transfers/conversions within the external structure
7. Knowledge transfers/conversions from external to internal structure
8. Knowledge transfers/conversions from internal to external structure
9. Knowledge transfers/conversions within internal structure
10. Maximise Value Creation – See the Whole

**Figure No. 8.2**




**The Ten Knowledge Strategy Issues are:**

**1. Knowledge Transfers/Conversions**

Between Individual Professionals Knowledge transfers/conversions between individuals are concerned with how to best enable the communication between employees within in the organisation and determine what types environments are most conducive to creativity. The strategic questions are: How can we improve the transfer of competence between people in the organisation? How can we improve the collaborative climate? The most important issues are probably concerning trust in the organisation. How willing are people to share their ideas and what they know?

Answers to such questions lead towards activities focused on trust building, enabling team activities, induction programs, job rotation, master/apprentice schemes, etc.

**Examples:** Oticon, the Danish hearing-aid manufacturer established in 1905, has re-designed whole work areas to create an atmosphere of openness, flexibility, creativity and sharing. The company emphasizes "live" interaction. Stand-up coffee bars encourage impromptu meetings, and dialogue rooms" with a table and chairs help employees relax while solving problems or sharing knowledge. Oticon even locked up elevators so there would be more "accidental" meetings in the stairwell. The company believes that paperwork hampers the exchange of information because it is slower and more formal than oral communication. The company therefore designated a "paper room," the only room where paper is "safe." Even electronic mail is discouraged in favor of face-to-face communication. These tactics have contributed towards live dialog becoming an integral part of Oticon's business, so much so that other forms of communication are almost non-existent.

Personnel rotation programs are common, and expose employees to expertise held locally and tacitly and are common. For instance, every executive including the CEO at Southwest Airlines spends at least one day every quarter as a baggage handler, ticket agent, or flight attendant. This "shop-floor" experience keeps the knowledge of the operation fresh in the minds of all employed. It also improves communication across all levels.

**2. Knowledge Transfers/conversions from Individuals to External Structure:**

Knowledge transfers/conversions from individuals to the external structure are concerned with how the organisation's employees transfer their knowledge to the outer world. The strategic question is: How can the organisation's employees improve the competence of customers, suppliers and other stakeholders? Answers to such questions lead towards activities focused on empowering the employees to help the customers learn about the products, getting rid of red tape, doing job rotation with customers, holding product seminars, providing customer education, etc.

**Examples:** Consultants at McKinsey, the US based consulting firm, are encouraged to spend time on publishing their research and methods in order to build the reputation of the firm. Baxter International markets healthcare products and has extended its offering to include service to hospitals. Baxter employees now mix drugs in intravenous solutions and act as brokers for other vendors.

**3. Knowledge Transfers/conversions from External Structure to Individuals:**

Employees learn a lot from customer, supplier and community feedback such as ideas, new experiences, feedback and new technical knowledge. Knowledge transfers/conversions from the external structure to individuals

are concerned with how the organisation's employees can learn from the external structure. Organisations tend to have procedures in place that capture such knowledge but they are scattered, not measured and hence do not systematically influence strategy formulation. The strategic question is: How can the organisation's customers, suppliers and other stakeholders improve the competence of the employees? Answers to such questions lead towards activities focused on creating and maintaining good personal relationships between the organisation's own people and the people outside the organisation.

**Examples:** Employees at Betz Laboratories in Trevose, Pennsylvania, frequently participates in its customers' quality management teams in order to gain a better understanding of, and even anticipate, customer needs. This knowledge is used to develop products that will boost customer sales. Betz measures value added from this knowledge by tracking its customers' return on investment, and its own employees receive awards for outstanding efforts to increase these returns.

**4. Knowledge Transfers/conversions from Competence to Internal Structure:**

Huge investments are currently being made in order to convert competence (often tacitly held) individual into data repositories. The idea is that information in such repositories will be shared with the whole organisation. Indeed, the marketers of database software have been so successful that many managers believe that buying a database is equal to "Knowledge Management". To focus one's investments on databases and document handling etc. will realise only a fraction of the value of a more strategic approach based on a knowledge-based theory of the firm.

The strategic question is: How can we improve the conversion of individually held competence to systems, tools and templates? Answers to this question lead towards activities focused tools, templates, process and systems so they can be shared more easily and efficiently. Examples systems for medical diagnostics, intranets, document handling systems, databases, etc.

The key to create value from database or intranet system is not the sophistication of the technology but on the climate in the firm and the level of involvement from all agents in the system. The US chemicals manufacturer Buckman Labs is well-known for nurturing a collaborative climate despite the fact that its 1,300 associates are spread all over the world. The company has been using electronic means for capturing experiences and information since 1987. It's new products to sales ratio went from ~25% to >35% when it began involving the customers in their intranet in 1994.

**5. Knowledge Transfers/conversions from Internal Structure to Individual Competence:**

This is the counterpart of the 4th strategy. Competence "captured in a system" is information and this information needs to be made available to other individuals in such a way that they improve their capacity to act; otherwise, the investment is a waste. IT systems can by definition only produce information. The key to value creation is whether the information generates competence. The strategic question is: How can we improve individuals' competence by using systems, tools and templates? Answers to such questions lead towards activities focused on improving the human-computer interface of systems, action-based learning processes, simulations and interactive e-learning environments.

Examples: IKEA, the Swedish furniture company, uses customised simulations for speeding up the learning of its warehouse employees.

The Copeland Corporation, a manufacturer of compressors, changed its entire manufacturing approach based on the results of a single demonstration effort, in which a multifunctional team designed a demonstration factory to manufacture a new product line. Experimentation, whether an ongoing program or a demonstration project, helps individuals move from superficial knowledge to a more basic understanding of its processes—from knowing about something to learning how and why.

**6. Knowledge Transfers/conversions within the External Structure:**

What do the customers tell each other about the services/products of a supplier? How are the products used? The conversations among the constituencies can have an enormous impact on the strategy of a company. Strategy formulation from a knowledge perspective adds a richer range of possible activities to traditional customer satisfaction surveys and one-way PR-activities. The company can support the competence growth of customers and influence how competence is transferred also between the stakeholders in the external structure. The strategic question is: How can we enable conversations among the customers, suppliers and other stakeholders to improve their competence to serve their customers? Answers to such questions lead towards activities focused on partnering and alliances, improving the image of the organisation and the brand equity of its products and services; improving the quality of the offering; conducting product seminars and alumni programs. Examples: Danish biomedical producer Novo actively engages in building local communities to improve the image of its products in its local community. Book publisher Berrett-Koehler runs seminars for its book buyers featuring its authors as speakers.

**7. Knowledge Transfers/conversions from External to Internal Structure**

Knowledge Transfers/conversions from External to Internal Structure are concerned with what knowledge the organisation can gain from the external world and how such new knowledge can be converted into action.

The strategic question is: How can competence from the customers, suppliers and other stakeholders improve the organisation's systems, tools & processes and products? Answers to such questions lead towards activities focused on empowering call centres to interpret customer complaints, creating alliances to generate ideas for new products, R&D alliances, etc.

**Example:** Frito-Lay, the US potato chips maker provides an interesting case of product differentiation of a commodity. The company uses its sales force to collect data about their customers. The data are analysed and fed back to their sales people empowering them with superior customer knowledge and competitive intelligence. Frito-Lay representatives not only use the information themselves, but they also give it away for "free" provided the shop buys their potato chips rather than their competitors'.

**8. Knowledge Transfers/conversions from Internal to External Structure**:
This is the counterpart of strategy 7. The strategic question is: How can the organisation's systems, tools & processes and products improve the competence of the customers, suppliers and other stakeholders? Answers to such questions lead towards activities focused on making the organisation's systems, tools & processes effective in servicing the customer, extranets, product tracking, help desks, business, etc.

**Examples:** Ernst & Young has created a tax and legal database, "Ernie", which allows its clients to tap into the data sources used also by its own consultants. 12 Ritz Carlton, the hotel chain renowned for its service, has installed a customer information database with global access. All staff are required to fill in cards with information from every personal encounter with a guest. These data plus guest profiles are stored and made available to staff in order to ensure personal treatment of all guests.

**9. Knowledge Transfers/conversions within Internal Structure**
The internal structure is the supporting backbone of the organisation. The strategic question is: How can the organisation's systems, tools & processes and products be effectively integrated? Answers to such questions lead towards activities focused on streamlining databases, building integrated IT systems, improving the office layout, etc.

**Example:** Again, this is a field dominated by Enterprise Systems and other company-wide IT solutions. Knowledge Curve, Pricewater house Cooper's intranet integrates several thousands of databases previously held individually or locally.

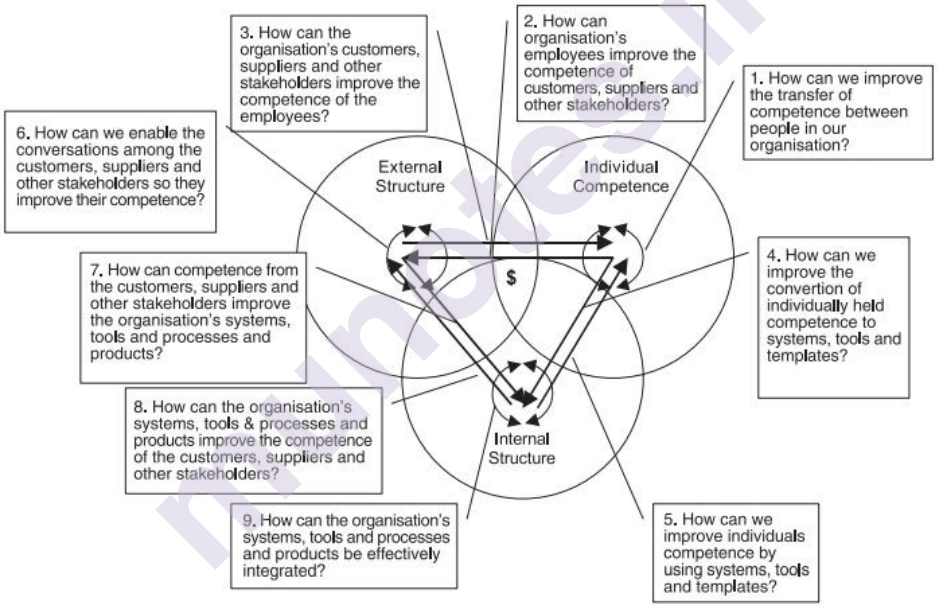
**10. Maximise Value Creation – See the Whole**:
The nine knowledge transfers/conversions exist in most organisations. However, they tend not to be coordinated in a coherent strategy, because management lack the full perspective that a knowledge-based theory may give them. Most organisations also have legacy systems and cultures that

block the leverage. Therefore, many of good initiatives go to waste or neutralize each other.

Investment in a sophisticated IT system for information sharing is for instance a waste of money if the organisation's climate is highly competitive – only junk will be shared. Reward systems that encourage individual competition will effectively block efforts to enhance knowledge sharing. Lack of standards and poor taxonomies reduce the value of document handling systems. A program for knowledge sharing with customers is neutralised by red tape protecting commercial secrets. Efforts to use ex-employees for building marketing relationships are useless if people leave the firm alienated or alumni programs are delegated to the administrative function. Data repositories do not improve individuals' capacity to act unless the databases are made highly interactive.

The Affärsvärlden case (see below) illustrates how important it is to integrate all activities in a strategic framework, so they leverage each other and do not neutralise investments made in other areas.

**Figure No. 8.3**




**Affärsvärlden's Knowledge-based Strategy :**

The competition between the two weekly Swedish business magazines Affärsvärlden (AFV) and Veckans affärer (VA) offer's a vivid illustration of the value of a knowledge-based strategy in publishing, one of the oldest industries on earth. There are substantial advantages of scale in the printing process, since loading the press with plates and paper and adjusting it represents a large fixed cost; after that, the marginal cost of printing the second copy is no more than maybe 10% of the average cost. Thus the larger the imprint, the lower the cost per page. The leverage is not as extreme as copying a CD, but not far from it.

The cost advantage enjoyed by a larger paper enables it to hire good journalists and maintain a higher overall level of editorial quality. This can

be a ticket to a virtuous circle of more readers who provide more resources, which enable better quality, which attracts more readers, etc.

Even if the smaller magazine keeps lower prices than the larger, the larger and (for the magazine) more profitable advertisers tend to prefer to place their ads in a large magazine, 14 because it gives them access to a larger audience. Publishers know that, once established, the largest newspaper or magazine in a market is a licence to print money.

AFV, being less than one tenth the size of VA was close to bankruptcy in 1977. The printing costs alone were 30% higher for AFV, even though its pages contained only half as much full-color print as VA´s. The journal's new owners in 1978 thus faced a formidable competitive barrier and had no alternative but to try a different strategy than VA. AFV adopted a more knowledge-based strategy.

**Affärsvärlden´s Knowledge Strategy:**
The knowledge strategy gave Affärsvärlden a distinct competitive advantage on the market for financial information. The strategy and some of the activities went against "common sense" in publishing, but its ultimate success made Affärsvärlden a "cult publication" in the 1980s and created a following among other journals and publishers in the country. A summary of Affärsvärlden´s knowledge strategy is found in Figure 4. below.

Sveiby (1994) identified two features that contributed most of the difference in margin during the period 1980 - 1993:

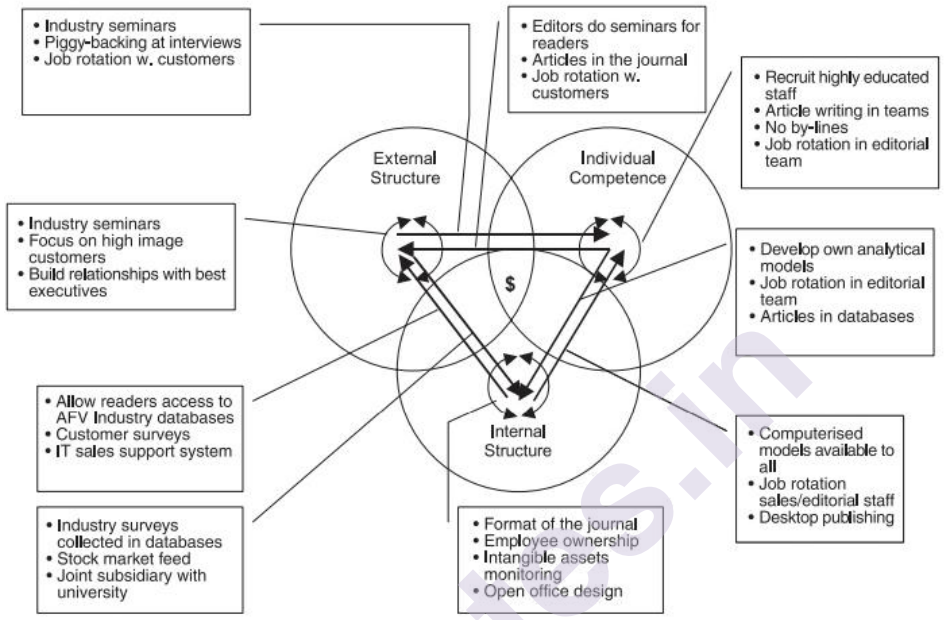
1. High editorial productivity. In 1983-84 the AFV journalists wrote twice as many pages as their colleagues on the VA staff (133 compared to VA´s 62). This difference in editorial productivity was sustained for 15 years. The knowledge-based strategy initiatives were:

    - Recruit highly educated staff. AFV journalists had access to more expertise in-house because they all had MBAs or higher, whereas VA's journalists rarely held such degrees. Higher education also gives competence in information processing.
    - Create Collaborative climate. No individual by-lines on the articles reduced the traditional competitive climate among journalists. Articles written by teams, "piggybacking" at interviews and master/apprentice model supported tacit knowledge transfer. Open office design (also in sales departments and for managers) supported informal information knowledge transfers/conversions.
    - Build flat organisation. Visible managers, employee ownership and profit sharing contributed to a shared vision and the collaborative climate.
    - Invest in new editorial technology. AFV was at least one year faster than the competitor VA, sometimes 2 years, in

implementing the new technologies that revolutionised publishing during the 1980s.

- Computerise analytical models. AFV's analysts were early in computerising their analytical models and basic number crunching. Computerisation freed up time to do more qualified analyses.

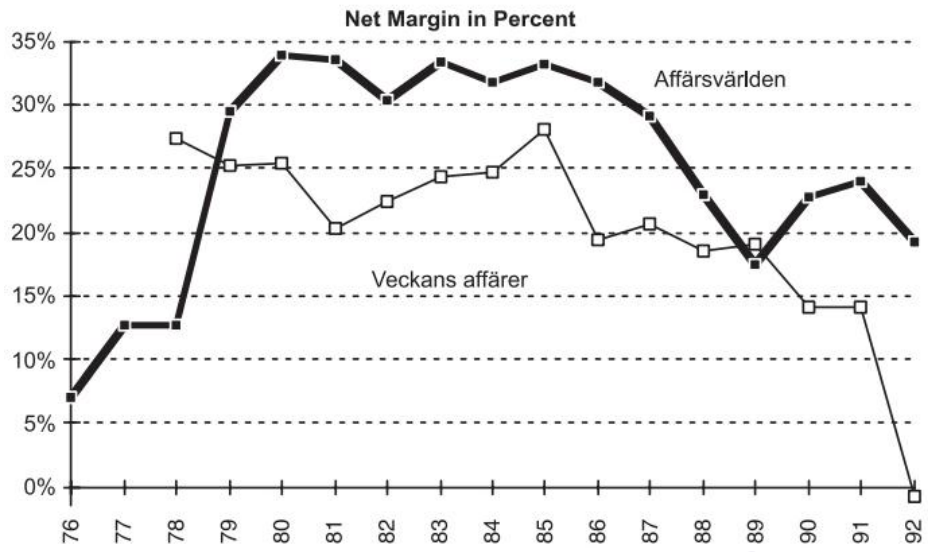
**Figure No. 8.4**




2. Low Staff turnover. The financial markets in the 1980s were exploding (such as the IT markets in the 1990s) and AFV's financial analysts were prime targets of the investment bankers. The ownership model and the collaborative culture were the strategic initiatives 15 that worked as "golden handcuffs" and kept the staff turnover at 5-7% throughout the whole period, while VA suffered at least twice the turnover.

AFV proved the value of its strategy by being more profitable than VA almost the whole period, see below, figure 8.5.

**Figure No. 8.5**




When the depression of the nineties hit the Swedish financial markets, both magazines came under heavy pressure. VA was hardest hit because its editorial concept involved high fixed costs. AFV had lower fixed costs and a much more flexible concept and was thus able to adjust rapidly by reducing the number of pages and cutting down its fixed costs. AFV continued to operate at a profit while VA went into the red.

Veckans Affärer´s problems caused its publishers to decide, with effect from March 1994, to split it into two journals: a smaller, cheaper weekly with a different concept and a more expensive monthly.

After 18 years (!) of single-minded head-on competition, AFV had forced the leader to move out.

## 8.4 TRANSACTION COST THEORY

The transaction cost approach to the theory of firm was created by Ronald Coase. Transaction Cost refers to the cost of providing for some good or service through the market rather than having it provided from within the firm.

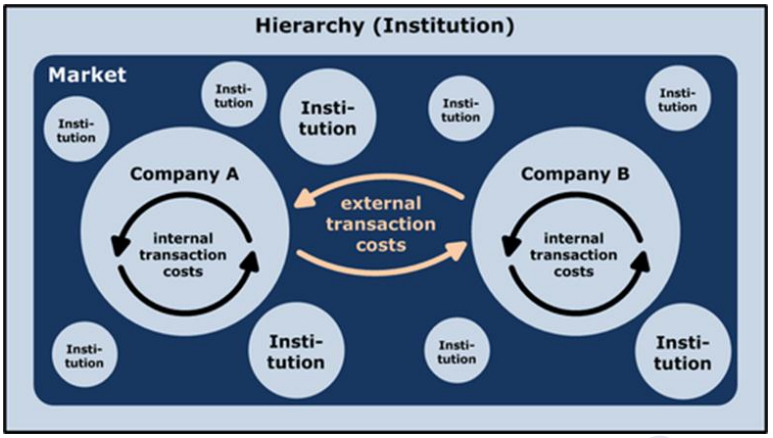
Which components should a manufacturing firm make in-house, which should it co-produce, and which should it outsource? Who should sit on the firm's board of directors? What is the right balance between debt and equity financing?

These questions may appear different on the surface, but they are all variations on the same theme: how should a complex contractual relationship be governed to avoid waste and to create transaction value?

Transaction Cost Economics (TCE) is one of the most established theories to address this fundamental question.

**Ronald Coase:**

**Figure No. 8.6**




The model shows institutions and market as a possible form of organization to coordinate economic transactions. When the external transaction costs are higher than the internal transaction costs, the company will grow. If the external transaction costs are lower than the internal transaction costs the company will be downsized by outsourcing.

According to Ronald Coase's essay The Nature of the Firm, people begin to organise their production in firms when the transaction cost of coordinating production through the market exchange, given imperfect information, is greater than within the firm.

Ronald Coase set out his transaction cost theory of the firm in 1937, making it one of the first (neo-classical) attempts to define the firm theoretically in relation to the market. One aspect of its 'neoclassicism' lies in presenting an explanation of the firm consistent with constant returns to scale, rather than relying on increasing returns to scale. Another is in defining a firm in a manner which is both realistic and compatible with the idea of substitution at the margin, so instruments of conventional economic analysis apply.

He notes that a firm's interactions with the market may not be under its control (for instance because of sales taxes), but its internal allocation of resources are: “Within a firm, … market transactions are eliminated and in place of the complicated market structure with exchange transactions is substituted the entrepreneur … who directs production.”

Coase begins from the standpoint that markets could in theory carry out all production, and that what needs to be explained is the existence of the firm, with its "distinguishing mark … [of] the supersession of the price mechanism." Coase identifies some reasons why firms might arise, and dismisses each as unimportant:

1. if some people prefer to work under direction and are prepared to pay for the privilege (but this is unlikely);
2. if some people prefer to direct others and are prepared to pay for this (but generally people are paid more to direct others);
3. if purchasers prefer goods produced by firms.

Instead, for Coase the main reason to establish a firm is to avoid some of the transaction costs of using the price mechanism. These include discovering relevant prices (which can be reduced but not eliminated by purchasing this information through specialists), as well as the costs of negotiating and writing enforceable contracts for each transaction (which can be large if there is uncertainty). Moreover, contracts in an uncertain world will necessarily be incomplete and have to be frequently re-negotiated. The costs of haggling about division of surplus, particularly if there is asymmetric information and asset specificity, may be considerable.

If a firm operated internally under the market system, many contracts would be required (for instance, even for procuring a pen or delivering a presentation). In contrast, a real firm has very few (though much more complex) contracts, such as defining a manager's power of direction over employees, in exchange for which the employee is paid. These kinds of contracts are drawn up in situations of uncertainty, in particular for relationships which last long periods of time. Such a situation runs counter to neo-classical economic theory. The neo-classical market is instantaneous, forbidding the development of extended agent-principal (employee-manager) relationships, of planning, and of trust. Coase concludes that “a firm is likely therefore to emerge in those cases where a very short-term contract would be unsatisfactory”, and that “it seems improbable that a firm would emerge without the existence of uncertainty”.

He notes that government measures relating to the market (sales taxes, rationing, price controls) tend to increase the size of firms, since firms internally would not be subject to such transaction costs. Thus, Coase defines the firm as "the system of relationships which comes into existence when the direction of resources is dependent on the entrepreneur." We can therefore think of a firm as getting larger or smaller based on whether the entrepreneur organises more or fewer transactions.

The question then arises of what determines the size of the firm; why does the entrepreneur organise the transactions he does, why no more or less? Since the reason for the firm's being is to have lower costs than the market, the upper limit on the firm's size is set by costs rising to the point where internalising an additional transaction equals the cost of making that transaction in the market. (At the lower limit, the firm's costs exceed the market's costs, and it does not come into existence.) In practice, diminishing returns to management contribute most to raising the costs of

organising a large firm, particularly in large firms with many different plants and differing internal transactions (such as a conglomerate), or if the relevant prices change frequently.

Coase concludes by saying that the size of the firm is dependent on the costs of using the price mechanism, and on the costs of organisation of other entrepreneurs. These two factors together determine how many products a firm produces and how much of each.

**Reconsiderations of transaction cost theory:**
According to Louis Putterman, most economists accept distinction between intra-firm and interfirm transaction but also that the two shade into each other; the extent of a firm is not simply defined by its capital stock. George Barclay Richardson for example, notes that a rigid distinction fails because of the existence of intermediate forms between firm and market such as inter-firm co-operation.

Klein asserts that "Economists now recognise that such a sharp distinction does not exist and that it is useful to consider also transactions occurring within the firm as representing market (contractual) relationships." The costs involved in such transactions that are within a firm or even between the firms are the transaction costs.

Ultimately, whether the firm constitutes a domain of bureaucratic direction that is shielded from market forces or simply "a legal fiction", "a nexus for a set of contracting relationships among individuals" is "a function of the completeness of markets and the ability of market forces to penetrate intra-firm relationships".

The Transactions Cost Theory of the Firm focuses on problems of asymmetric information involved in transactions. The firm, according to this theory, comes into existence because it successfully minimises 'make' inputs costs (through vertical integration) and 'buy' inputs costs (using available markets). The more specific the inputs that the firm needs are the more likely it is that it would produce them internally and/or acquire them through joint ventures and alliances. The weakness of this theory is that it does not take into consideration agency costs or firm evolution, neither does it explain how vertical integration should take place in the face of investments in human assets with unobservable value, that cannot be transferred.

The Principal–Agent Theory of the Firm extends the neoclassical theory by adding agents to the firm. The theory is concerned with friction due to asymmetric information between owners of firms and their stakeholders or managers and employees; the friction between agent and principal, requires precise measurement of agent performance and the engineering of incentive mechanisms. The weaknesses of the theory are many: it is difficult to engineer incentive mechanisms, it relies on complicated

incomplete contracts (borderline unenforceable), it ignores transaction costs (both external and internal), and it does not allow for firm evolution. Coase's answer was that firms exist because they reduce transaction costs, such as search and information costs, bargaining costs, keeping trade secrets, and policing and enforcement costs.

**Further Additions:**
Ronald H. Coase, in 1937, was the first to highlight the importance of understanding the costs of transacting, but TCE as a formal theory started in earnest in the late 1960s and early 1970s as an attempt to understand and to make empirical predictions about vertical integration ("the make-or-buy decision"). TCE has become one of the most influential management theories, addressing not only the scale and scope of the firm but also many aspects of its internal workings, most notably corporate governance and organization design. TCE is therefore not only a theory of the firm, but also a theory of management and of governance.

At its foundation, TCE is a theory of organizational efficiency: how should a complex transaction be structured and governed so as to minimize waste? The efficiency objective calls for identifying the comparatively better organizational arrangement, the alternative that best matches the key features of the transaction. For example, a complex, risky, and recurring transaction may be very expensive to manage through a buyer-supplier contract; internalizing the transaction through vertical integration offers an economically more efficient approach than market exchange.

TCE seeks to describe and to understand two kinds of heterogeneity. The first kind is the diversity of transactions: what are the relevant dimensions with respect to which transactions differ from one another? The second kind is the diversity of organizations: what are the relevant alternatives in which organizational responses to transaction governance differ from one another? The ultimate objective in TCE is to understand discriminating alignment: which organizational response offers the feasible least-cost solution to govern a given transaction? Understanding discriminating alignment is also the main source of prescription derived from TCE.

The key points to be made when examining the logic and applicability of TCE are:

(1) The first phenomenon TCE sought to address was vertical integration, sometimes dubbed "the canonical TCE case." But TCE has broader applicability to the examination of complex transactions and contracts more generally.

(2) TCE could be described as a constructive stakeholder theory where the primary objective is to ensure efficient transactions and avoidance of waste. TCE shares many features with contemporary stakeholder management principles.

(3) TCE offers a useful contrast and counterpoint to other organization theories, such as competence- and power-based theories of the firm. These other theories, of course, symmetrically inform TCE.

Consider a situation in which two parties interested in a complex exchange of goods or services are trying to determine the best way of organizing the transaction. Both want to ensure their interests are being served, and both want to avoid unnecessary costs, delays, and wasted effort. Both also realize that all transactions involve risk but that unnecessary risks must be avoided. How are they to proceed with organizing the transaction? What kind of a contract will they strike?

In a resource-constrained world, seeking economic efficiency is always not only relevant but also common sense: if there are several alternative ways of conducting a business transaction, why not choose the one that consumes less resources? At the same time, in a world where work is complex, the future is uncertain, and both rationality of decision makers and availability of information are constrained, choosing the best among feasible alternatives requires effort, skill, foresight, and prudence.

At the most general level, Transaction Cost Economics (TCE) is a theory of how business transactions are structured in challenging decision environments. TCE is chiefly concerned with transactions that are complex in that they are recurring, subject to uncertainty, and involve commitments that are difficult to reverse without significant economic loss.

The more general question underpinning the make-or-buy decision pertains to governance of contractual relationships. Williamson elaborates: "Transaction cost economics holds that economizing on transaction costs is mainly responsible for the choice of one form of capitalist organization over another. It thereupon applies this hypothesis to a wide range of phenomena—vertical integration, vertical restrictions, labor organization, corporate governance, finance, regulation (and deregulation), conglomerate organization, technology transfer, and, more generally, to any issue that can be posed directly or indirectly as a contracting problem. As it turns out, large numbers of problems that on first examination do not appear to be of a contracting kind turn out to have an underlying contracting structure." In this section, we explore in detail this general contracting structure is and how it can be applied.

Let us return to the general premise that TCE starts at trying to specify how transactions differ. According to TCE, the three dimensions that merit attention are frequency, uncertainty, and specificity. All three should be thought of as characteristics of a contractual exchange relationship between two exchange parties; the principal unit of analysis in TCE is indeed the individual transaction.

(1) Frequency refers to the volume of transactions between the two exchange parties. Contractual relationships are always associated with

a cost, and with larger volumes (i.e., recurring transactions), costs of specialized governance structures can be justified, for instance (Williamson, 1985,).

(2) Uncertainty refers to the contracting parties' limited ability to predict environmental changes and one another's behavior under unforeseen circumstances. The two exchange parties always have interests that are only partially overlapping, and disagreements are a source of cost. In complex exchange relationships, it is simply impossible to write a complete contract that covers all possible contingencies. TCE works out of the assumption that contracts are incomplete.

(3) Specificity refers to specialized investments made by one party, or both parties, to enable the exchange.

Of the three dimensions, specificity deserves closer attention. For example, the supplier may build a sub-assembly plant that is co-located with the customer's final assembly plant. The economic value this sub-assembly plant generates would suffer greatly should the exchange relationship terminate. More generally, specificity takes many different forms: site specificity (e.g., an electric plant), physical asset specificity (e.g., specialized tools), and human asset specificity (e.g., firm-specific knowledge). Importantly, specificity gives rise to dependency, which may be either unilateral or bilateral. In many situations, even though the actual investment may appear on the balance sheet of just one of the transacting parties (e.g., investing in the sub-assembly plant), some kind of mutual dependency tends to develop over time. If the customer were to terminate the contract with the supplier who made the specific investment, it would either have to make the same investment itself, or alternatively, convince another supplier to do so. Of course, a dependency relationship is always at least somewhat asymmetric, and purely unilateral dependency tends to be rare in situations that involve specificity. In the complete absence of specificity, markets are competitive in the sense that no buyer is dependent on a specific supplier, or vice versa.

Commitment to specificity can create a situation in which one party to the transaction may see a possibility to take advantage of the other party. Indeed, such economic "holdup problems" (Goldberg)sometimes occur in practice. The position taken by TCE is that taking advantage of one's exchange partner by engaging in opportunistic behavior is both ill-advised and myopic. Williamson labeled opportunism "a very primitive response" that has an adverse consequence on transaction efficiency. Transacting parties who are about to commit to specificity should be wiser than that. A better option is to engage in farsighted contracting that is based on both giving and receiving credible commitments to support the exchange relationship. Exchanging credible commitments is, among other things, aimed at avoiding a potential holdup problem developing into an actual problem.

A simple transaction has low frequency, low uncertainty, and low specificity. Such transactions can be efficiently handled through a market transaction between a supplier and a buyer. For example, purchase of the carton of milk from the grocery store the transaction is routine in that it has little uncertainty, low asset specificity, and virtually no risk associated with it: therefore, the transaction is most efficiently handled through a straightforward market exchange. TCE provides an explanation for why simple transactions are organized as market transactions between a buyer and a seller, but provides insight particularly in the context of complex transactions that involve high degrees of specificity.

A supplier of make-and-model-specific components or sub-assemblies to a final automobile assembly plant is a good example. Applying the TCE logic, Monteverde and Teece predicted that automakers would be more likely to make in-house components that required greater make-and-model-specific applications engineering. In contrast, components whose specifications are known ex ante immediately become candidates for competitive bidding and outsourcing because the transaction costs are presumed to be comparatively lower. Monteverde and Teece maintained that the problem with the supplier's acquiring of transaction-specific know-how is a higher supplier switching cost on the part of the buyer. If the relationship were to terminate, the buyer would need to find another supplier who would need to develop the same transaction-specific know-how. This know-how would likely be difficult to transfer from the previous supplier.

The same line of thinking can be applied to many other decisions made within and across firms. Consider a company's mix of debt and equity financing. The choice is, of course, between alternative financial instruments, but also between alternative governance structures. The decision of debt versus equity financing is thus analogous to the vertical integration decision, where the key factor to consider is again specificity. Assets of low specificity are more effectively financed through debt. Because low-specificity assets are by definition redeployable, the lender will be covered in case the borrower defaults on the loan; no additional contractual safeguards are needed to manage risks. Consequently, the cost of transacting is relatively low. This is why car rental companies, for example, are able to rely on debt financing and various leasing arrangements for their vehicle fleet.

For a nuclear power plant, in contrast, debt financing is generally not feasible. Who is willing to accept highly specific, nonredeployable property as collateral? If the firm wanted to use debt to finance such assets, it would either have to pay a very high interest on the capital or to try to reduce asset specificity to enhance redeployability. The former would be prohibitively costly, indeed, most banks will probably not lend at any price. The latter may be either impossible or, at least, have significant adverse consequences such as increased production costs and lower quality. A better option is to finance high-specificity assets using a

governance mode where the financier does not receive a collateral-backed fixed interest but is instead made a recipient of the earnings that the specialized assets create. This solution, of course, leads to equity financing.

The choice of debt versus equity financing has a number of important organizational ramifications that pertain to monitoring and control. In firms financed largely by equity, the role of the board of directors is crucial in securing the rights of the providers of equity, the residual claimants. This economic safeguard is needed because there is no contract between the firm and the providers of equity that protects the interests of the latter. In a debt-financed firm, in contrast, the rights of the financier are stipulated in the loan agreement and in corporate law, effectively eliminating the need for additional safeguards. More generally, firms that rely on debt financing tend to organize based on formalization (rule-following); discretion is more dominant in equity-financed firms. Again, TCE emphasizes that financing decisions should also be considered contracting problems—with important managerial and organizational implications.

The objective of the early TCE scholars was to develop a theory that could be used as a source of empirical predictions about firm boundaries, management, and governance. Why would an automaker produce some components in-house and outsource others? Why would a firm lean toward equity as opposed to debt financing? Why would a public corporation appoint an employee representative on its board of directors?

Reflecting upon nearly four decades of empirical research, Williamson concluded that "TCE is an empirical success story" in that it had achieved its main objectives of producing testable empirical predictions.

Comparison and Criticism

Coase's main purpose was to explain why economic activity was organized within firms, since the works of Williamson, the TCT has shifted away from Coase's initial and more general treatment to concerns with issues such as appropriation, ownership, alignment of incentives, and self-interest.

Williamson state explicitly that the core methodological properties are

(1) the transaction is the basic unit of analysis

(2) the human agents are subject to bounded rationality and self-interest

(3) the critical dimensions for describing transactions are frequency, uncertainty, and transaction specific investments

(4) economizing on transaction costs is the principle factor that explains viable modes of contracting and

(5) assessing transaction cost differences is a comparative institutional exercise.

Alchian and Demetsz (1972) examined team production, information costs, and economic organization –contrasting transaction and production costs. Spence (1975), on the internal economics of the organization, suggested that resource allocation processes that are internalized are those which are not efficiently carried out in a decentralized manner (that is to say, where equilibria are inefficient).

**Criticism:**

Notwithstanding the tremendous impact of TCT of management research in the last two decades, TCT has been subjected to multiple criticisms. The TCT arguments have not remained unchallenged.

The most common criticism is that the central assumptions of TCT are flawed. For example, the assumption of opportunism has been criticized for ignoring the contextual grounding of human actions and therefore presenting an undersocialized view of human motivation and oversocialized view of institutional control. Williamson responded to such criticisms by re-stating that in his model, opportunism or bounded rationality may differ from person to person much as personality or intelligence do, but when transaction costs change they do so because of changes in the environment, not in the person.

Ghoshal and Moran attacked the validity of TCT on the grounds that the opportunism with guile is bad for practice. TCT is normative or prescriptive theory and if opportunism with guile assumption is taken seriously by managers there will be negative consequences for organizations. Application of TCT will increase the occurrence of opportunism rather than decreasing it.

Ghoshal and Moran also criticized TCT for failing to point out how opportunism is reduced through alternative governance structures. Jones argued that the problem with TCT is Williamson's description of the determinants of opportunism; and that there is a difference between the propensity to behave opportunistically (a behavioral trait) and the psychological state of opportunism. The same uncertainty condition that may lead some individuals to behave opportunistically it may lead others to trust. Under certain circumstances trust or cooperation may be the most rational and efficient self-interested behavior. The propensity to trust or opportunism as a state is a much more realistic assumption about human behavior given uncertainty.

Williamson treats environmental uncertainty as a threat that must be managed through the governance structure that allows managers to economize on transaction costs. Jones (1998) adopted a positive or entrepreneurial view and argued that bounded rationality and uncertainty are not problems to be managed and overcome, but rather are opportunities to be taken advantage of.

The TCT has been further criticized as only looking into two relative extremes methods of facilitating transactions that do not really exist. The critics argued that the market versus hierarchy dichotomy is somewhat misleading since many transactions are actually carried out through a hybrid governance form. But, Williamson stated that the distributions of transactions would be a "bell-shaped" normal distribution if discrete transaction would be located at the one extreme (market), highly centralized and hierarchical transactions on the other, and hybrid transactions (franchising, joint ventures, and other forms of nonstandard contracting) in between.

A major critic to TCT is its tautological nature. Eccles claimed that Williamson failed to operationalize the measurements of transaction costs and there is a tautological flavor in his arguments. Eccles argued that "ex-post arguments can usually be found that any given structure economized on transaction costs by simply defining these costs in a necessary way. When this cannot be done, the argument can be made that the existing structure is a 'mistake' and will eventually be replaced by one that does economize on these costs". According to Dow, the simple comparison of transaction costs under different governance structure is meaningless because the governance structure used to manage a transaction changes the nature of a transaction.

Jones noted that transaction costs appear on both the left and the right-hand sides of the causality equation, which is one of the typical attributes of tautologies. Although Williamson distinguished ex ante costs (such as negotiation costs) from ex post costs (such as costs associated with contractual failures), it is hard to find any costs that are not transaction costs.

Finally, TCT is criticized for failing to explain the alternative forms of organization and a lot of other organizational phenomena. However, TCT does not claim itself as panacea for everything; it only attempts to explain a portion of the organizational phenomena: why and under what conditions transactions are organized in certain ways (Coase, Williamson). At best, TCT deals with relative efficiency question. Therefore, while deserving a prominent place among the theories in organization, TCT can and should not be used exclusively to explain organization phenomena.

**Conclusion:**

Transaction cost theory or transaction cost economics has become an increasingly important anchor for the analysis of a wide range of strategic and organizational issues of considerable importance to firms.

To conclude, it is undeniable the merit of the TCT for examining firms 'choices, namely those regarding where to set the boundaries of the firms. Or, as some scholars put it, choices regarding what they do and what they do not. But, to Jones, TCT is "not a flawed transplant from economics but a valuable addition and refinement to organizational theory that has taken

the analysis of organizational issues and the theory of the firm to a new level of sophistication".

We observe that the influence of TCT is enormous in the management disciplines and albeit we see that other concepts and views are emerging – such as the resource-, knowledge-, capabilities-based view - the TCT will likely maintain its influence in the discipline.

## 8.5 SUMMARY

In this way we study the existence and very purpose of the firm. Resources, of all types play an important role in the existence and growth of the firm. Similarly, the knowledge and the extent of it helps in expansion of the firm.

## 8.6 QUESTIONS

Q1. Write a note on existence of firm.

Q2. Write a note on the horizontal and vertical boundaries of firm.

Q3. Explain the resource-based theory of firm.

Q4. Explain in details the knowledge-based theory of firm

Q5. Write an explanatory note on the transaction-based theory of firm.

Q6. Explain transaction cost theory of firm.

## 8.7 REFERENCES


- Arrow, K. J. (1971). Essays in the theory of risk bearing. New York: North Holland.
- Gravelle H. and Rees R.(2004) : Microeconomics., 3rd Edition, Pearson Edition Ltd, New Delhi.
- Gibbons R. A Primer in Game Theory, Harvester-Wheatsheaf, 1992
- A. Koutsoyiannis : Modern Microeconomics
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.
- Varian H (2000): Intermediate Microeconomics: A Modern Approach, 8th Edition, W.W.Norton and Company Varian: Microeconomic Analysis, Third Edition
- Salvatore D. (2003), Microeconomics: Theory and Applications, Oxford University Press, New Delhi.
- Williamson, O. E. (1988). Corporate finance and corporate governance. Journal of Finance, 43, 567–591.


*****