UNIT – I

Money in Economics: Definition, Types, Functions, Characteristics, Importance and

Evils | Economics

Evolution of Money:

As barter system was an inconvenient method of exchange, people were compelled to select some commodity which was most commonly accepted in that area as a medium of exchange. Thus, a large variety of goods came to be used as money; gradually the most attractive metals, like gold, silver, etc., were adopted as money almost everywhere.

Money has now taken the place of all these commodities. Later coins were replaced or supplemented by paper currency for the reasons of economy and convenience. The bank cheques, drafts and promissory notes came into use in addition of currency to serve as the most important type of money. However, today each country has its own monetary system and the money of one is not usually acceptable outside its borders.

In fact, this is one of the reasons which makes international trade different from internal trade. Money was not invented overnight. The development of money was rather slow. It is the result of a process of evolution through several hundred years.

The different types of money indicate the different stages of the development of money. Wheat, corn, tobacco, skins, beads, gold, etc. Even live animals served as a medium of exchange at different times in different parts of the world. Rulers in all lands found that making coins is a profitable business and took it into their own hands.

Meaning and Definitions of Money:

The word “money” is believed to originate from a temple of ‘Juno’, located on Capitoline, one of Rome’s seven hills. In the ancient world Juno was often associated with money. The temple of Juno Moneta at Rome was the place where the mint of Ancient Rome was located.

The name “Juno” may derive from the Etruscan goddess Uni (which means “the one”, “unique”, “unit”, “union”, “united”) and “Moneta” either from the Latin word “monere” (remind, warn or instruct) or the Greek word “moneres” (alone, unique).

Now-a-days everybody recognizes money but usually does not know how to define money. Money has been defined differently by different economists. While some economist like WALKER has defined money in terms of the functions, while others like KEYNES, COLE, ROBERTSON, etc., have emphasized on the general acceptability aspect of it.

To serve as money, the definition of money should be comprehensive enough to cover all the essential functions that money performs in the economy. Before we arrive at the most suitable definition, it is essential to study a few definitions of money as given by some eminent economists.

Definitions of Money:

Money is one such concept which is very difficult to be restricted to some well-defined set of words. It is very easy to understand but difficult to define. Still, a large number of economists have given variety of definitions, some definitions are too extensive while others are too narrow. Various economists like Prof. Walker, Robertson, Seligman, etc., have used different characteristics for defining it.

Legal Tender Money and fiduciary Money:

Legal tender money is issued by the monetary authority of a country. It has legal sanction of the Government. Every individual is bound to accept legal tender money in exchange for goods and services, and in the discharge of debts.

Legal tender money is of two kinds:

(a) Limited legal tender, and

(b) Unlimited legal tender.

Fiduciary optional money is non-legal tender money as it is generally accepted by the people in final payments. It comprises credit instruments like cheques, drafts, bills of exchange, etc. Acceptance of optional money depends upon the will of a person.

Stages in the Evolution of Money:

(i) Animal Money:

In ancient India, Go-Dhan (cow wealth) was accepted as form of money. Similarly, in the fourth century B.C., the Roman State had officially recognized cow and sheep as money to collect fine and taxes.

(ii) Commodity Money:

The second stage in the evolution of money is the introduction of commodity money. Commodity money is that money whose value comes from a commodity, out of which it is made. The commodities that were used as medium of exchange included cowrie shells, bows and arrows, gold, silver, food grains, large stones, decorated belts, cigarettes, copper, etc.

However, the commodity money had various drawbacks such as there could be no standardization of value for money, lacks the property of portability and indivisibility. Therefore this form of money became an unsuitable medium of exchange.

(iii) Coinage:

The next step is coinage. This is just like a commodity money but the commodity is the metal that the money is made of. Thus, it can be seen that commodity money is of two types i.e., metallic and non-metallic.

When the use of money was not so very extensive, copper could do the job but when the number of transactions increased gradually, silver and then gold was used as a main metal for money and coins of small denominations were prepared either of copper or of silver.

b) Unlimited legal tender.

Fiduciary optional money is non-legal tender money as it is generally accepted by the people in final payments. It comprises credit instruments like cheques, drafts, bills of exchange, etc. Acceptance of optional money depends upon the will of a person.

Stages in the Evolution of Money:

(i) Animal Money:

In ancient India, Go-Dhan (cow wealth) was accepted as form of money. Similarly, in the fourth century B.C., the Roman State had officially recognized cow and sheep as money to collect fine and taxes.

(ii) Commodity Money:

The second stage in the evolution of money is the introduction of commodity money. Commodity money is that money whose value comes from a commodity, out of which it is made. The commodities that were used as medium of exchange included cowrie shells, bows and arrows, gold, silver, food grains, large stones, decorated belts, cigarettes, copper, etc.

However, the commodity money had various drawbacks such as there could be no standardization of value for money, lacks the property of portability and indivisibility. Therefore this form of money became an unsuitable medium of exchange.

(iii) Coinage:

The next step is coinage. This is just like a commodity money but the commodity is the metal that the money is made of. Thus, it can be seen that commodity money is of two types i.e., metallic and non-metallic.

When the use of money was not so very extensive, copper could do the job but when the number of transactions increased gradually, silver and then gold was used as a main metal for money and coins of small denominations were prepared either of copper or of silver.

(iv) Paper Money:

The next important stage in the evolution of money is the paper money which replaced the metallic money. The transfer of sum of money in terms of metallic money was both inconvenient and risky. Therefore, written documents were used as temporary substitutes for money. Any person could deposit money with a wealthy merchant or a goldsmith and get a receipt for the deposit.

These receipts and documents were not actual money but temporary substitutes of money. This marked the development of paper money. These paper notes gradually took the form of currency notes.

(v) Bank Money:

As the volume of transactions increased, paper money started becoming inconvenient because of time involved in its counting and space required for its safe-keeping. This led to the introduction of bank money (or credit money).

Bank money implies demand deposits with banks which are withdraw able through cheques, drafts, etc. Cheques are widely accepted these days particularly for business transactions. Debit and credit cards also fall under this category.

Characteristics of Money:

1. General Acceptability:

Money is accepted by all as a medium of exchange. Thus, it has general acceptability. No one denies to accept money as a medium of exchange. People do not hesitate to accept it as standard of payment.

2. Measure of Value:

Value of any good or service can easily be measured in terms of money. It is accepted as a measure of value.

3. Active Agent:

Money is an active agent of an economic system. In modern economy, money is required in every commercial process. Process of production cannot start without the participation of money.

4. Liquid Assets:

Money is highly liquid asset. It can easily be converted in goods and services. Debt, stock and bills, etc., are the other liquid assets but the liquidity of money is highest than the other liquid assets. One has to first get to convert other liquid assets into money, then it can be converted in desired goods or services, while money can directly be converted.

5. Money is a Means and not an End:

The word money is means to acquire things desired. Money itself cannot be used to satisfy. It is indirectly used to get any goods or services to satisfy human wants.

6. Voluntary Acceptability:

Money is voluntarily accepted by people. There is no requirement to get legal approval. People always wish to hold money.

7. Government Control:

Reserve Bank of India and Govt, of India have an authority to issue currency which is accepted as a form of money in India. No other authority can issue currency notes. Thus, the government keeps control over the money supply in the country.

Classification of Money:

Money assumes so many forms in real life that it is difficult to identify what constitutes money and what not. Different economists have classified money in different forms.

The more important classifications of money are as follows:

(i) Actual Money and Money of Account:

Actual money is that which actually circulates in the economy. It is used as a medium of exchange for goods and services in a country. For example, paper notes of different denominations and coins in actual circulation in India constitute the actual money. Money of account is that form of money in terms of which the accounts of a country are maintained and transactions made.

For example, rupee is the money of account in India. Generally, actual money and money of account are the same for a country; however, sometimes actual money may be different from the money of account. For example, rupee and paise is the money of account in India. In real practice, however, one paisa coin is nowhere visible.

(ii) Commodity Money and Representative Money:

Commodity money is made up of a certain metal and its face value is equal to its intrinsic value. It is also referred to as full-bodied money. Representative money, on the other hand, is generally made either of cheap metals or paper notes. The intrinsic value of the representative money is less than its face value. Currency notes and coins are good examples of representative money in India. Representative money may or may not be converted into full-bodied money.

(iii) Money and Near-Money:

Money is anything that possesses 100 per cent liquidity. Liquidity is the quality of being immediately and always exchangeable in full value for money. Near-money refers to those objects which can be held with little loss of liquidity. For example, National Savings Deposits, Building Society Deposits and other similar deposits are not money because they are not generally acceptable in paying debt; these, however, could be easily and quickly exchanged for money without any loss or with minimum loss.

(iv) Metallic Money and Paper Money:
This classification is based upon the content of a unit of money. Money made of some metal like gold and silver is called metallic money. On the other hand, money made of paper, such as currency notes, is called paper money.

Metallic money is sub-classified into:
(a) Standard Money, and

(b) Token Money.
Standard money is one whose intrinsic value is equal to its face value. It is made up of some precious metal and has free coinage. Token money is that form of money whose face value is higher than its intrinsic value. Indian rupee coin is an example of token money. Paper money comprises bank notes and government notes which circulate without difficulty.

Paper money is classified into following parts:
(a) Representative paper money, which is 100 per cent backed and is fully redeemable in some precious metal.

(b) Convertible paper money, which can be converted into standard coins at the option of the holder. It is not fully backed by precious metals.

v) Credit Money:
It is also known as bank money. This consists of deposits of the people held with the banks, which are payable on demand by the depositors. Cheques, drafts, bills of exchange, etc., are examples of credit money.

Modern Forms of Money:
1. Currency:
The currency is a country's unit of exchange issued by their government or central bank whose value is the basis for trade. Currency includes both metallic money (coins) and paper money that is in public circulation.
(a) Metallic Money:
Metallic money refers to the coins which are used for small transactions. Coins are most often issued by the government. Examples of coins are 50 paise coins, and 1, 2, 5 and 10 rupee coins.
(b) Paper Money:
It refers to paper notes and used for large transactions. Each currency note carries the legend, 'I promise to pay the bearer the sum of 50/100 rupees'

depending on the value of note. The currency notes are duly signed by the Governor of RBI.

Simply, the meaning of legend is that it can be converted into other notes or coins of equal value. Examples of currency notes are 1, 2, 5, 10, 20, 50, 100, 500 and 2000 rupee notes.

2. Deposit Money or Bank Money:

It refers to money deposited by people in the bank on the basis of which cheques can be drawn. Customers of the bank deposit coins and currency notes in the bank for safe-keeping, money transferring and also to get interest on the deposited money.

This money is recorded as credit to the account of the bank's customer which can be withdrawn by him on his/her wish by cheques. Cheques are widely accepted these days because transfer of money through cheques is convenient.

3. Legal Tender Money (Force Tender):

Legal tender money is the currency which has got legal sanction or approval by the government. It means that the individual is bound to accept it in exchange for goods and services; it cannot be refused in settlement of payments of any kind.

Both coins and currency notes are legal tender. They have the backing of government. They serve as money on the fiat (order) of the government. But a person can legally refuse to accept payment through cheques because there is no guarantee that a cheque will be honored by the bank in case of insufficient deposits with it.

Currency is the most common form of legal tender. It is anything which when offered in payment extinguishes the debt. Thus, personal cheques, credit cards, debit cards and similar non-cash methods of payment are not usually legal tenders.

Coins and notes are usually defined as a legal tender. The Indian Rupee is also legal tender in Bhutan but Bhutanese Ngultrum is not legal tender in India.

4. Near Money:

It is a term used for those which are not cash but highly liquid assets and can easily be converted into cash on short notice such as bank deposits and treasury bills. It does not function as a medium of exchange in everyday purchases of goods and services.

5. Electronic Money:

Electronic money (also known as e-money, electronic cash, electronic currency, digital money, digital cash or digital currency) involves computer networks to perform financial transactions electronically. Electronic Funds Transfer (EFT) and direct deposit are examples of electronic money. The financial institutions transfer the money from one bank account to another by means of computers and communication links. A country wide computer network would monitor the credits and debits of all individuals, firms, and government as transactions take place in the economy.

It exchange funds every day without the physical movement of any paper money. This would eliminate the use of cheques and reduce the need for currency.

6. Fiat Money:

Fiat money is any money whose value is determined by legal means. The term fiat currency and fiat money relate to types of currency or money whose usefulness results not from any intrinsic value or guarantee that it can be converted into gold or another currency but from a government's order (fiat) that it must be accepted as a means of payment.

A distinction between money and currency may be made here. The term 'currency' includes only metallic coins and paper notes which are legal tender and are in actual circulation in the country. The term 'money' however includes not only currency in circulation but also credit instruments. In other words, we may say that all currency is money but all money is not currency.

Importance of Money:

Money plays a significant role in modern economy. It has an active role in economic activities.

Importance of money in an economy can be discussed as below:

1. Money and Production:

Money helps in various ways in the process of production. Money can help producers to decide, plan, execute and manage the production activities. Moreover, the existence of money helps the producers to assess the quality and quantity of demand of a consumer.

2. Money and Consumption:

Money has a great importance in consumption. Consumers with the help of the money can easily decide, what they want and how much. They have a ready command over the goods and services. Moreover, they can postpone their demands, if required.

3. Money and Distribution:

Money has made it possible to distribute the reward accurately and conveniently among the various factors of production. The reward can be distributed in terms of wages, rent, interest and profit in the form of money.

4. Removal of the Difficulties of Barter:

There were some difficulties attached to the barter system of exchange, i.e., lack of double coincidence of wants, problem of measurement of value, problem of future payment, etc. Invention of money has overcome all the difficulties of barter system. There is no need to find double coincidence of wants and value can be measured easily in terms of money.

5. Money and Capital Formation:

Money is essential to facilitate capital formation. Savings of people can be mobilized in the form of money and these mobilized savings can be invested in more profitable ventures. Financial institutions are the part of this process. They mobilize the savings and channelize them in productive process.

6. Money and Public Finance:

Public finance deals with the income and expenditure of the government. Government receives its income in the form of money through taxes and other means and make expenditures in development and administrative processes.

7. External Trade:

Money has facilitated trade not only inside the country but also outside countries. With the use of money, goods and services can easily and rapidly be exchanged. Though in external trade foreign currencies are used in receipts and payments but they are exchanged with the help of domestic currencies.

8. Money and Economic Development:

Supply of money in a country affects its economic development. If the money supply is more, then it may lead to inflationary situation in the economy which may hamper growth. Similarly, if the supply of money is lesser than what is required then there will be shortage of liquidity which will lead to lesser investments and hence lesser employment.

Value of Money:

The value of money means all is related with its exchange value. Apart from exchange value of money it has no other independent value. In other words, the money is always related with its exchange value. As we know the eye whether of human person or animal does not have its own light, similarly the eye can see only with either by artificial or natural light. In the same way, the value of money can be judged or perceived only when it is related with its power of purchase.

In the words of Crowther "The value of money is what is will buy." In other words the value of money depends on its purchasing power. In this connection the other definition of Robertson may also be referred. As per this definition— **"The value of money means the amount or things in general which will be given in exchange for a unit of money."**

The solution of such problem has been found out on the following three consecutions:

(1) Wholesale Value:

Whatever value becomes prevalent in the wholesale market is usually taken as wholesale value. So, the wholesale value is easy to be found out because the value of money usually is displayed on this very base. This is called the wholesale value of the money.

(2) Retail Value:

The value prevalent in the retail market is called as retail value. But the retail value may be perceived separately on different places. This means the retail value will remain constant. The calculation of the retail value is always different from one place to another and as such the base of retail price is difficult in comparison to wholesale price.

(3) Labour Value:

In order to make payment the money among the labourers the value prevalent in such a market is usually called the value of labour. Now the value of labour will never be constant and it will also vary from place to place. So, it cannot be accepted as bases of value.

Evils of Money:

Money is not an unmixed blessing. It is said that money is a good servant but a bad master.

Several evils of money are said to be:

(i) Economic Instability:

Several economists are of the opinion that money is responsible for economic instability in capitalist economies. In the absence of money, saving was equal to investment. Those who saved also invested. But in a monetized economy, saving is done by certain people and investment by some other people. Hence, saving and investment need not be equal. When saving in an economy exceeds investment, then national income, output and employment decrease and economy falls into depression.

On the other hand, when investment exceeds saving, then national income, output and employment increase and that leads to prosperity. But if the process of money creation and investment continues beyond the point of full employment, inflationary pressures will be created. Thus inequality between saving and investment are known to be main cause of economic fluctuations.

The main evil of money lies in its liability of being over-issued in the case of inconvertible paper money. The over-issue of money may lead to hyper-inflation. Excessive rise in prices brings suffering to the consuming public and fixed income earners. It encourages speculation and inhibits productive enterprises. It adversely affects distribution of income and wealth in the community so that the gulf between the rich and poor increases.

(ii) Economic Inequalities:

Money is a very convenience tool for accumulating wealth and of the exploitation of the poor by the rich. It has created an increasing gulf between the ‘haves’ and the ‘have-nots. The misery and degradation of the poor is, thus, in no small measure due to the existence of money.

(iii) Moral Depravity:
Money has weakened the moral fiber of man. The evils to be found in the affluent society are only too obvious. The rich monopolizes all the social evils like corruption, the wine and the woman. In this case, money has proved to be a soul-killing weapon.

(iv) Medium of Exploitation:
Prominent socialist like Marx and Lenin condemned money but it helps the rich to exploit the poor. When the communists came to power in Russia, they tried to abolish money. But they soon realized that to run a modern economy without money was impossible. All economic activity has to be based on monetary calculations. Accordingly, money is fully and firmly established in all Socialists States. Money performs several functions like facilitating optimum allocation of the country's resources, functions as a medium of exchange and a measure of value, guides economic activity and is essential for facilitating distribution of national income.

<u>Role of money in capitalistic socialistic and mixed economics</u>

- The money has huge role in capitalistic socialistic and mixed economics. The capitalistic economy means the the individuals and the business entity focus on **earning profit based on their investment.**
- The primary role of money is **investing.** The socialistic economy depends on the government.
- The government concentrates on production based on investment.
- The mixed economy is the mixture of the **private and the governmental enterprise.** The money from government managed by private institution.
    - ROLE OF MONEY IN CAPITALIST ECONOMY
    - What to Produce: This is the first function of prices, it will help us decide what to produce. Resources in any economy are limited or scarce, so they must be allocated according to in relation to the total output. Hence prime mechanism will help dictate what goods to be produced.
    - How much to produce: The next question that needs answering is the quantity of production required. The resources are limited, so the quantity will depend on the preferences and requirements of the society and the relative prices of the product.
    - How to Produce: Then the decision must be taken about the method of production. Say for example you are producing cotton textiles. Should

you employ a labor-intensive method or an automated method? This will depend on the availability and prices of the factors of production (labor, capital etc.)

- For whom to Produce: An economy cannot satisfy the needs and wants of every person. So price mechanism will decide how to distribute the total output among its citizens.
- Provisions for Economic Growth: An economy cannot use up all of its limited resources. It has to make provisions for the future, so the economy can grow. Otherwise, the economy, income levels, output etc will stagnate and may even decline. So the level of savings and investment must be decided.
- Role of Price Mechanism in Capitalist, Socialist and Mixed Economy
-
- Role of Price Mechanism in Socialist Economy
-
- In a socialist economy, the decisions of what, how and for whom to produce is not dependent on market forces or price mechanism, these decisions are taken by the Central Planning Authority.
-
- So while price mechanism does play a role in a socialist economy, it is a very minimal role. It is used to ensure the disposal of stock that has accumulated in the economy. Since the allocation of resources is planned by the authorities, price mechanism will have no say here. And there is no profit motive in socialism, so again price mechanism has no role in the area.

- Role of Price Mechanism in Mixed Economy

- Here price mechanism will pay an equal role along with the planning authority. Especially in the private sector of such an economy, the price mechanism along with the competing forces helps the economy with an allocation of resources and other such efficient decisions.
- The decision of what to produce in the private sector will depend on market forces, but in the public sector that decision will fall on the central planning authority. Prices will be fixed by the authority on profit-price policy or no-profit no-loss policy. How to produce goods

will also depend both on price mechanism and government interaction.

Types of Monetary Standards: Metallic and Paper Standard | Economics

A. Metallic Standard:

Under metallic standard, the monetary unit is determined in terms of some metal like gold, silver, etc. Standard coins are made out of the metal. Standard coins are full-bodied legal tender and their value is equal to their intrinsic metallic worth. The important thing to note is that to be on a metallic standard a country must keep – (a) its monetary unit at a constant value in terms of the selected metal, and (b) its various types of money convertible into the selected metal at constant values.

Metallic standard may be of two types:

. Monometallism

2. Bimetallism.

1. Monometallism:

Monometallism refers to the monetary system in which the monetary unit is made up or convertible to only one metal. Under monometallic standard, only one metal is used as standard money whose market value is fixed in terms of a given quantity and quality of the metal.

Features of Monometallism:

Essential features of monometallic standard are given below:

(i) Standard coins are defined in terms of only one metal.

(ii) These coins are accepted as unlimited legal tender in the discharge of day-to-day obligations.

(iii) There is free coinage (i.e., manufacture of coins) of the metal.

(iv) There are no restrictions on the export and import of metal to be used as money.

(v) Paper money also circulates, but it is convertible into standard metallic coins.

Types of Monometallism:

Monometallism can be of two types:

a. Silver Standard:

Under silver standard, the monetary unit is defined in terms of silver. The standard coins are made of silver and are of a fixed weight and fineness in terms of silver. They are unlimited tender. There is no restriction on the import and export of silver. The silver standard remained in force in many countries for a long period.

India remained on silver standard from 1835 to 1893. During this period, Rupee was the standard coin and its weight was fixed at 180 grains and fineness 11/12. The coinage of the Rupee was free and people can get their silver converted into coins at the mint. Similarly, silver coins could be melted into bullion.

Silver standard lacks universal recognition as compared to gold standard. There is greater instability of both internal and external values of money under silver standard because silver price fluctuates more than that of gold. Thus, as far as the metal is concerned, gold is preferred to silver in most of the countries.

b. Gold Standard:

Gold standard is the most popular form of monometallic standard; the monetary unit is expressed in terms of gold. The standard coins possess a fixed weight and fineness of gold. The gold standard remained widely accepted in most of the countries of the world during the last quarter of the 19th century and the first quarter of the 20th century.

The U.K. was the first country to adopt the gold standard in 1816. She was also the first to abandon this standard in 1931. Germany adopted the gold standard in 1873, France in 1878 and the U.S.A. in 1900. Gradually, gold standard disappeared from different countries and finally it was completely abandoned by the world by 1936.

Gold standard is the most popular form of monometallic standard. Under gold standard, the monetary unit is expressed in terms of gold. The standard coins possess a fixed weight and fineness of gold. The gold standard remained widely accepted in most of the countries of the world during the last quarter of the 19th century and the first quarter of the 20th century. The U.K. was the first country to adopt the gold standard in 1816.

She was also dying first to abandon this standard in 1931. Germany adopted the gold standard in 1873, France in 1878 and the U.S.A. in 1900. Gradually, gold standard disappeared from different countries and finally it was completely abandoned by the world by 1936.

Gold standard has been defined differently by different monetary economists. According to D.H. Robertson, "Gold standard is a state of affairs in which a country keeps the value of its monetary unit and the value of a defined weight of gold at equality with one another." According to Coulborn, "The gold standard is an arrangement whereby the chief piece of money of a country is exchangeable with a fixed quantity of gold of a specific quality."

In the words of Kemmerer, "a gold standard is a monetary system in which the unit of value, in which price and wages are customarily expressed, and in which the debts are usually contracted, consists of the value of a fixed quantity of gold in an essentially free gold market."

Merits of Monometallism:

Monometallic standard has the following advantages:

i. Simplicity:

Since only one metal is used as a standard of value, monometallism is simple to operate and easy to understand.

ii. Public Confidence:

Since the standard money is made of a precious metal (gold or silver), it inspires public confidence.

iii. Promotes Foreign Trade:

Monometallism facilitates and promotes foreign trade. Gold or silver standard is easily acceptable as an international means of payment.

iv. Avoids Gresham's Law:

Monometallism avoids the operation of Gresham's law. According to this law, when both good as well as bad money exist in the economy, bad money tends to drive out of circulation good money.

v. Self-Operative:

It makes the supply of money self-operative. If there is surplus money supply, the value of money will fall and the people will start converting coins into metal. This will wipe out the surplus money, thus creating a balance.

Demerits of Monometallism:

The following are the demerits of monometallism:

. Costly Standard:

It is a costly standard and all countries, particularly the poor countries, cannot afford to adopt it.

ii. Lacks Elasticity:

Monometallism lacks elasticity. Money supply depends upon the metallic reserves. Thus, money supply cannot be changed in accordance with the requirements of the economy.

iii. Retards Economic Growth:

Economic growth requires expansion of money supply to meet the increasing needs of the economy. But, under monometallism, scarcity of metal may create scarcity of money supply which, in turn, may hinder economic growth.

iv. Lacks Price Stability:

Since the price of the metal cannot remain perfectly stable, the value of money (or the internal price level) under monometallism lacks stability.

2. Bimetallism:

Bimetallism is a monetary system which attempts to base the currency on two metals. According to Chandler, “A bimetallic or double standard is one in which the monetary unit and all types of a nation’s money are kept at constant value in terms of gold and also in terms of silver.” Under bimetallism two metallic standards operate simultaneously.

Two types of standard coins from two different metals (say gold and silver) are minted. Both the types of standard coins become unlimited legal tender and a fixed ratio of exchange based on mixed ratio of exchange based on mint parity is prescribed for them. Provisions for unlimited purchase, sale and redeem-ability are extended to both metals.

Features of Bimetallism:

(i) A bimetallic standard is based on two metals; it is the simultaneous maintenance of both gold and silver standards.

(ii) There is free and unlimited coinage of both metals.

(iii) The mint ratio of the values of gold and silver at the mint is fixed by the government.

(iv) Two types of standard coins (i.e., gold coins and silver coins) are in circulation at the same time.

(v) Both the coins are full-bodied coins. In other words, the face value and the intrinsic value of both the coins are equal

(vi) Both the coins are unlimited legal tenders. They are also convertible into each other.

(vii) There is free import and export of both the metals.

Merits of Bimetallism:

The merits of bimetallism are discussed below:

i. Convenient Full-Bodied Currency:

Bimetallism provides convenient full-bodied coins for both large and small transactions. It provides portable gold money for large transactions and convenient silver money for smaller payments. This argument has, however, lost its force now when credit money has developed.

ii. Price Stability:

Under this monetary system, the shortage of one metal can be offset by increasing the output of the other metal. Consequently, stability in the prices of both the metals and hence, in the internal prices can be ensured.

iii. Exchange Rate Stability:

Bimetallism ensures stability of exchange rate. As long as gold and silver are stabilised in terms of each other, the currencies of all countries with fixed values in gold or in silver would exchange for each other at nearly constant rates.

iv. Sufficient Money Supply:

Under bimetallism, sufficient money supply is assured to meet the trade requirements of the economy. Since there is no question of both metals becoming scarce simultaneously, money supply is more elastic under this system.

v. Maintenance of Bank Reserves:

Under bimetallism, the maintenance of bank reserves becomes easy and economical. Under this system, both gold and silver coins are standard coins and unlimited tender. Therefore, it is easy for the banks to keep their cash reserves either in gold coins or in silver coins or in both.

vi. Low Interest Rates:

Since, under bimetallism, money is made of two metals, its supply is generally more than its demand. As a result, the interest rates decline. Banks can extend loans at cheaper rates. This would increase investment and hence production in the economy.

vii. Stimulates Foreign Trade:

Bimetallism stimulates international trade in two ways, – (a) A country on bimetallism can have trade relations with both gold standard and silver standard countries, (b) There are no restrictions on imports and exports due to the free inflow of both types of coins.

Demerits of Bimetallism:

Bimetallism has the following demerits:

i. Operation of Gresham's Law:

Bimetallism in a single country is a temporary and not workable monetary standard due to the operation of Gresham's law. According to this law, when there is a disparity between the mint parity rate and the market rate of exchange of the two metals, bad money or the over-valued metal at the mint (whose mint price exceeds market price) tends to drive out of circulation good money or under-valued metal at the mint (whose market price exceeds mint price).

Thus, ultimately, single metal money (monometallism) will remain in practice. Thus, national bimetallism is only a temporary phenomenon. Only international bimetallism can prove permanent and practicable.

ii. Inequality between Mint and Market Rates:

Bimetallism can operate successfully only if the equality between the market rate and the mint rate can be maintained. But, in practice, it is difficult to maintain equality between the two rates, particularly when one metal is oversupplied than the other.

iii. No Price Stability:

The argument that bimetallism ensures internal price stability and there will be an automatic adjustment between supply and demand for money is illusionary. There can be a possibility of both the metals to become scarce.

iv. Payment Difficulties:

Bimetallism leads to difficult situation in the settlement of transactions when one party insists on payment in terms of a particular type of coins.

v. Encourages Speculative Activity:

It encourages speculative activity in the two metals when their prices fluctuate in the market.

vi. No Stimulus to Foreign Trade:

International trade is stimulated if all the countries adopt bimetallism. But, this is a rare possibility in the present circumstances.

vii. Costly Monetary Standard:

Bimetallism is a costly monetary standard and all nations, particularly the poor nations, cannot afford to adopt it.

Gresham's Law:

Gresham's law in its simple form states that when good and bad money are together in circulation as legal tender, bad money tends to drive good money out of circulation. This implies that less valuable money tends to replace more valuable money in circulation.

This law was enunciated by Sir Thomas Gresham who was the financial adviser to Queen Elizabeth I in the 16th century in England. Gresham was, however, not the first to develop this law, but it became associated with his name after he explained a problem faced by the Queen. With a view to reform the currency system, the Queen tried to replace bad coins of the previous regime by issuing new full-weighted coins.

But to her surprise, as soon as new coins were circulated, they disappeared and the old debased coins continued to remain in circulation. She sought the advice of Sir Thomas Gresham, who provided his explanation in the form of the law which states- "Bad money tends to drive out of circulation good money."

The theoretical explanation of this law is in terms of the divergence of the market rate of exchange of the two currencies from mint rate. If the mint rate (i.e., the official rate of exchange between two types of money) differs from the market rate of exchange between the two types of money, then the over-valued money at the mint will tend to drive the under-valued money out of circulation.

Suppose under bimetallism, one gold coin exchanges for 10 silver coins, i.e., the official rate of exchange or the mint rate is 1:10. Now, if the market rate is 1:12, then gold is under-valued and silver is over-valued at the mint rate (i.e. the market rate of gold exceeds the mint rate and the market rate of silver is less, than its mint rate). In this case, gold will become good money and silver a bad money. The bad money (silver) will drive out good money (gold) from circulation.

Operation of the Law:

When both good and bad money together are in circulation as legal tender, good money disappears in three ways:

i. Good Money is Hoarded:

When both good and bad money circulate simultaneously, people have the tendency to hoard good money and use bad money for making payment.

ii. Good Money is Melted:

Since both good coins and bad coins are in circulation and have the same value, people prefer to melt good coins to convert them into ornaments or other items of art.

iii. Good Money is Exported:

In payments to the foreign countries, gold coins are accepted by weight and not by counting. Thus, it would be profitable to pay to the foreigners in terms of new full-weight coins rather than old and light-weight coins.

Gresham's Law in General Form:

Gresham's law, in its original form, applies only to debased coins of monometallic system (i.e., gold standard).

But, the law can, however, be extended to all forms of monetary standards:

1. Under Monometallism:

Under monometallism (for example gold standard), the old and worn out coins are regarded as bad coins and full-weight coins are considered as good coins. According to Gresham's law, the old and worn out coins drive new and full- weight coins out of circulations.

2. Under Bimetallism:

Under bimetallism (generally a system of gold and silver coins), coins of overvalued metals are considered bad money and coins of under-valued metal as good money. Thus, according to Gresham's law, the over-valued coins will drive under-valued coins out of circulation.

3. Under Paper Standard:

Under paper standard, if both standard coins of superior metal and inconvertible paper notes are in circulation, the metallic coins will be good money and paper notes will be bad money. Thus, paper notes will drive out standard coins from circulation.

Thus, Gresham's law is a general law which can be applicable in different forms of monetary standards. Marshall presented a generalized version of the law – "Gresham's law is that an inferior currency, if not limited in amount, will drive out the superior currency."

Limitations of the Law:

Gresham's law will operate if the following necessary conditions are satisfied.

In the absence of these conditions the law will fail to apply:

i. Usefulness of Good Money:

An important condition for Gresham's law is that the intrinsically more valuable money (i.e., good money) must also be more valuable in other uses than it is as money in circulation. The failure of this condition to apply explains why the coin currency today remains in circulation as fairly as paper currency despite its higher intrinsic value.

ii. Fixed Parity Ratio:

The applicability of the law requires that the intrinsically more valuable money must be relatively fixed by law in its parity with money. The law will not hold where one money becomes intrinsically more valuable than another money (at the old parity) if the parity changes.

iii. Sufficient Money Supply:

The law will operate only if both good money and bad money are in circulation and the total money supply is more than the actual monetary requirements of the economy.

iv. Sufficient Supply of Bad Money:

The applicability of the law requires that there should be sufficient bad money in circulation to meet the transactions requirement of the people. If there is scarcity of bad money, both good and bad money will remain in circulation and the law will not operate.

v. Contents of Pure Metal:

The law will not operate if the contents of pure metal in coins are less than that in the old ones.

vi. Acceptability of Bad Money:

The law will operate if people are prepared to accept bad money in transactions.

vii. Distinction between Good Money and Bad Money:

The law assumes that people can distinguish between bad money and good money.

viii. Development of Banking Habit:

The law applies in the absence of banking habits. Development of banking habits among the people tends to discourage hoarding and thus restricts the operation of Gresham's law.

ix. Convertibility:

The law also does not operate if the country is on inconvertible paper standard.

B. Paper Standard:

Paper standard refers to a monetary standard in which inconvertible paper money circulates as unlimited legal tender. Under paper money standard, although the standard money is made of paper, both currency and coins serve as standard money for purpose of payment. No gold reserves are required either to back domestic paper currency or to facilitate foreign payments.

The paper standard is known as managed standard because the quantity of money in circulation is controlled and managed by the monetary authority with a view to maintain stability in prices and incomes within the country. It is also called fiat standard because paper money is inconvertible in gold and still regarded as full legal tender. After the general breakdown of gold standard in 1931, almost all the countries of the world shifted to the paper standard.

Features of Paper Standard:

The paper standard has the following features:

(i) Paper money (paper notes and token coins) circulates as standard money and accepted as unlimited legal tender in the discharge of obligations.

(ii) The unit of money is not defined in terms of commodity.

(iii) The commodity value (or intrinsic value) of the circulating money is particularly nil.

(iv) Paper money is not convertible in any commodity or gold.

(v) The purchasing power of the monetary unit is not kept at par with any commodity (say gold).

(vi) Paper standard is national in character. There is no link between the different paper currency systems.

(vii) The foreign rate of exchange is determined on the basis of the parity of purchasing powers of the currencies of different countries.

Merits of Paper Standard:

Various merits of paper standard are described below:

1. Economical:

Since under paper standard no gold coins are in circulation and no gold reserves are required to back paper notes, it is the most economical form of monetary standard. Even the poor countries can adopt it without any difficulty.

2. Proper Use of Gold:

Wastage of gold is avoided and this precious metal becomes available for industrial, art and ornamental purpose.

3. Elastic Money Supply:

Since paper money is not linked with any metal, the government or the monetary authority can easily change the money supply to meet the industrial and trade requirements of the economy.

4. Ensures Full Employment and Economic Growth:

Under paper standard, the government of a country is free to determine its monetary policy. It regulates its money in such a way that ensures fall employment of the productive resources and promotes economic growth.

5. Avoids Deflation:

Under paper standard, a country avoids deflationary fall in prices and incomes which is the direct consequence of gold export. Such type of situation arises under gold standard when a participating country experiences adverse balance of payments. This results in the outflow of gold and contraction of money supply.

6. Useful during Emergency:

Paper currency is very useful in times of war when large funds are needed to finance war. It is also best suited to the less developed countries like India. To these economies, it makes available large amounts of financial resources through deficit financing for carrying out developmental schemes.

7. Internal Price Stability:

Under this system, the monetary authority of a country can establish stability in the domestic price level by regulating money supply in accordance with the changing requirements of the economy.

8. Regulation of Exchange Rate:

Paper standard provides more effective and automatic regulation of exchange rate, whereas, under gold standard, the exchange rate is absolutely fixed. Whenever, exchange rate fluctuates as a result of disequilibrium between demand and supply forces, the paper standard works on imports and exports and restores equilibrium. It allows the forces of demand and supply to operate freely to establish equilibrium.

Demerits of Paper Standard:

The paper standard suffers from the following defects:

1. Exchange Instability:

Since the currency has no link with any metal under paper currency, there are wide fluctuations in the foreign exchange rates. This adversely affects the country's international trade. Exchanging instability arises whenever external prices move more than domestic prices.

2. Internal Price Instability:

Even the commonly claimed advantage of paper standard, i.e., domestic price stability, may not be achieved in reality. In fact, the countries now on paper standard experience such violent fluctuations in internal prices as they experienced under gold standard before.

3. Dangers of Inflation:

Paper standard has a definite bias towards inflation because there is always a possibility of over- issue of currency. The government under paper standard generally has a tendency to use managed currency to cover up its budget deficit. This results in inflationary rise in prices with all its evil effects.

4. Dangers of Mismanagement:

Paper currency system can serve the country only if it is properly and efficiently managed. Even the minor mistake in the management of paper currency can bring such disastrous result that cannot be conceived of in any other form of monetary standard. If the government issues little more or little less currency than what is required for maintaining price stability, it may lead to cumulative inflation or cumulative deflation.

5. Limited Freedom:

In the present world of economic dependence, it is almost impossible for a particular country to isolate itself and remain unaffected from the international economic fluctuations simply by adopting paper standard.

6. Absence of Automatic Working:

The paper standard does not function automatically. To make it work properly, the government has to interfere from time to time.

Principles of Note Issue:

At present, all the countries of the world have adopted paper standard.

In fact this standard has proved a boon to the modern monetary system. The central bank of a country, which plays an important role in the paper standard, is assigned the job of note issue.

A good note issue system should possess the following qualities:

(a) It should inspire public confidence, and, for this, it must be based on sufficient reserves of gold and silver.

(b) It should be elastic in the sense that money supply can be increased or decreased in accordance with the needs of the country.

(c) It should be automatic and secure.

To ensure a good note issue system, two principles of note issue have been advocated:

(1) Currency principle and

(2) Banking principle.

1. Currency Principle:

The currency principle is advocated by the ‘currency school’ comprising Robert Torrens, Lord Overstone, G. W. Norman and William Ward. Currency principle is based on the assumption that a sound system of note issue should command the greatest public confidence. This requires that the note issue should be backed by 100 per cent gold or silver reserves. Or in other words, paper currency should be fully convertible into gold or silver.

Thus, according to the currency principle, the supply of paper currency is subjected to the availability of metallic reserves and varies directly with the variations in the metallic reserves.

Merits:

The currency principle has the following advantages:

(i) Since, according to this principle, the paper currency is fully convertible into gold and silver, it inspires maximum confidence of the public.

(ii) There is no danger of note issue of the paper currency leading to the inflationary pressures,

(iii) It makes the paper currency system automatic and leaves nothing to the will of the monetary authority.

Demerits:

The currency principle has the following drawbacks:

(i) The currency principle makes the monetary system inelastic because it does not allow the monetary authority to expand the money supply according to the needs of the country.

(ii) It requires full backing of gold reserves for note issue. Thus, it makes the monetary system expensive and uneconomical.

(iii) This principle is not practical for all countries because gold and silver are unevenly distributed among countries.

2. Banking Principle:

The banking principle is advocated by the ‘banking school’, the important members of which are Thomas Tooke, John Fullarton James, Wilson and J.W. Gilbart. The banking principle is based on the assumption that the common man is not much interested in getting his currency notes converted into gold or silver.

Therefore 100 per cent metallic reserves may not be necessary against note issue. It is sufficient to keep only a certain percentage of total paper currency in the form of gold or silver reserves. The banking principle of note issue is derived from the practice of the commercial banks to keep only a certain proportion of cash reserves against their total deposits.

Merits:

The following are the merits of banking principle:

(i) The banking principle renders note issue system elastic. The monetary authority can change the supply of currency according to the needs of the economy.

(ii) Since the banking principle does not require 100 percent metallic backing against the note issue, it is the most economic principle and thus can be followed by both rich and poor countries.

Demerits:

The banking principle has the following demerits:

(i) The banking principle is inflationary in nature, because it involves the danger of over-issue of paper currency.

(ii) The monetary system based on the banking principle does not command public confidence because the system is not fully backed by metallic reserves.

Conclusion:

The main conclusion regarding the two principle of note issue is:

(i) Both the currency principle and the banking principle fail to satisfy all the, requirements of a good note issue system. The currency principle ensures security and public confidence, but it lacks elasticity, economy and practicability. On the contrary, the banking principle provides elasticity and economy to the note issue system, but it suffers from the drawbacks of over-issue and loss of public confidence.

(ii) Despite the incompleteness of both the principles, the banking principle, rather than the currency principle, has been preferred and widely accepted in the modern times mainly because of its emphasis on the qualities of elasticity, economy and practicability of the note issue system. The quality of convertibility, which is basic to the currency principle, is no longer considered as necessary requirement for a good note issue system.

Methods of Note Issue:

Different countries have adopted various methods of note-issue in different periods.

Important methods of note-issue are discussed below:

1. Simple Deposit System:

Under the simple deposit system, the paper currency notes are fully backed by the reserves of gold or silver or both. This system is based on the currency principle of note issue. This method involves no danger of over-issue of currency and commands maximum degree of public confidence. But, this system has never been practised because it is very costly and has no elasticity of money supply.

2. Fixed Fiduciary System:

Under the fixed fiduciary system, the central bank is authorised to issue only a fixed amount of currency notes against government securities. All notes issued in excess of this limit should be fully backed by gold and silver reserves. Fiduciary issue means the issue of currency notes without the backing of gold and silver. This system was first introduced in England under the Bank Charter Act of 1844 and still prevails there. India followed this system between 1862 to 1920.

Merits:

Fixed fiduciary system has the following advantages:

(i) It ensure convertibility of currency notes.

(ii) It inspires public confidence since the government guarantees the convertibility of notes.

(iii) There is no danger of over-issue of paper notes because barring a certain portion, the entire note issue is backed by gold reserves.

Demerits:

The main disadvantages of the fixed fiduciary system are:

(i) It makes the monetary system less elastic. In times of economic crises, money supply cannot be increased without keeping additional gold in reserve.

(ii) It is a costly system which requires sufficient gold reserves. Poor countries cannot afford to adopt it.

(iii) It is inconvenient method because whenever gold reserves fall, the central bank has to reduce the supply of currency which greatly disturbs the functioning of the economy.

3. Proportional Reserves System:

Under the proportional reserve system, certain proportion of currency notes (40%) are backed by gold and silver reserves and the remaining part of the note issue by approved securities. India adopted this method on the recommendation of Wilton Young Commission.

According to the Reserve Bank of India Act 1933, not less than 40 per cent of the total assets of the Issue Department should consist of gold bullion, gold coins and foreign securities, with the additional provision that gold coins and gold bullion were not at any time to be less than Rs. 40 crores. The proportional reserve system was later replaced by the minimum reserve system by the Reserve Bank of India (Amendment) Act, 1956.

Merits:

The proportional reserve system has the following advantages:

(i) It guarantees convertibility of paper currency.

(ii) It ensures elasticity in the monetary system; the monetary authority can issue paper currency much more than that warranted by reserves.

(iii) It is economical and can be easily adopted by the poor countries.

Demerits:

The proportional reserves system suffers from the following defects:

(i) Under this system, it is easy to expand currency but very difficult to reduce it. The reduction of currency has deflationary effects in the economy.

(ii) There is wastage of gold because large amount of gold lies in the reserve and cannot be put to productive use.

(iii) The convertibility of paper notes is not real. In practice, high denomination notes are converted into low denomination notes and not into coins.

4. Minimum Fiduciary System:

Under the minimum fiduciary system, the minimum reserves of gold against note issue that the authority is required to maintain are fixed by law. Against these minimum reserves, the monetary authority can issue as much paper currency as is considered necessary for the economy. There is no upper limit fixed for the issue of currency.

Minimum fiduciary system is based upon two considerations:

(a) The central bank feels that there should not be any restriction on the note issue, especially when the demand for currency is fast increasing,

(b) In the modern age, when bank deposits assume greater significance as an important constituent of money supply, the convertibility of notes into gold need not be bothered about.

The minimum fiduciary system, if ably managed, can prove very useful for developing countries. It can make available enormous resources for financing developmental schemes. Similarly, during inflation, the monetary authority can control the money supply. This system has been in force in India since 1957. The Reserve Bank of India is required to keep minimum reserves of Rs. 200 crores of which not less than Rs. 115 crores must be kept in the form of gold.

Merits:

The minimum reserve system has the following advantages:

(i) The system is economical because the entire note issue need not be backed by metallic reserves. Only a minimum reserve is to be maintained.

(ii) It renders elasticity to the monetary system. After maintaining the minimum reserves, the monetary authority can issue any amount of currency that it feels necessary.

Demerits:

The minimum reserve system has the following drawbacks:

(i) Since, under this system, no additional reserves are required for increasing the supply of currency, there is always a tendency towards the over-issue of currency, and hence an inherent danger of inflationary pressures.

(ii) Since the system provides no convertibility of currency notes into gold, it lacks public confidence.

5. Maximum Reserve System:

Under this system, the government fixes the maximum limit upto which the monetary authority can issue notes without the backing of metallic reserves. The monetary authority cannot issue notes beyond this limit. The maximum limit is not rigid and may be revised from time to time according to the changing needs of the economy.

This system was followed by France and England upto 1928 and 1939 respectively. Under this system, the Central bank is given the power to determine the maximum limit and thus an element of elasticity is introduced in the system of note issue. The system, however involves the dangers of over-issue and loss of public confidence when the maximum limit is raised and additional currency is circulated without the backing of metallic reserves.

Conclusion:

The following conclusions emerge from the discussion of various methods of note issue:

(i) The analysis of relative merits and demerits of various methods of note issue makes it difficult to identify any one method as an ideal method.

(ii) A good method of note issue must possess the qualities of economy, elasticity and public confidence without being inflationary in effect.

(iii) Convertibility of currency notes into some precious metal is no longer considered an important requirement because in modern times currency notes are accepted on their own merit.

(iv) Keeping in view these considerations, minimum fiduciary system can prove to be a better method, if managed ably and sincerely.

Ideal Monetary System:

An ideal monetary standard should be able to achieve the twin objectives of – (a) growth and full employment with reasonable price stability within the country, and (b) smooth flow of goods, services and capital at the international level. Such an ideal monetary system requires wise blending of both paper and gold standards. This blending will provide the advantages of both the standards, with none of their disadvantages.

In modern times, the establishment of International Monetary Fund (IMF) and the International Bank of Reconstruction and Development (IBRD) has been designed to give the ideal monetary system a practical shape. These institutions have been able – (a) to make the paper standard function efficiently both internally and internationally by removing its various defects; and (b) to operate international affairs quite successfully, thus promoting trade and cooperation among the nations.

UNIT - I

MONETARY ECONOMICS

1 Introduction

Monetary economics is the economics of the money supply, prices and interest rates, and their repercussions on the economy. It focuses on the monetary and other financial markets, the determination of the interest rate, the extent to which these influence the behavior of the economic units and the implications of that influence in the macroeconomic context. It also studies the formulation of monetary policy, usually by the central bank or "the monetary authority," with respect to the supply of money and manipulation of interest rates, in terms both of what is actually done and what would be optimal.

In a monetary economy, virtually all exchanges of commodities among distinct economic agents are against money, rather than against labor, commodities or bonds, and virtually all loans are made in money and not in commodities, so that almost all market transactions in a modern monetary economy involve money.[1] Therefore, few aspects of a monetary economy are totally divorced from the role of money and the efficiency of its provision and usage, and the scope of monetary economics is a very wide one.

Monetary economics has both a microeconomics and a macroeconomics part. In addition, the formulation of monetary policy and central bank behavior – or that of "the monetary authority," often a euphemism for the central banking system of the country[2] – is an extremely important topic which can be treated as a distinct one in its own right, or covered under the microeconomics or macroeconomics presentation of monetary economics.

Microeconomics part of monetary economics

The microeconomics part of monetary economics focuses on the study of the demand and supply of money and their equilibrium. No study of monetary economics can be even minimally adequate without a study of the behavior of those financial institutions whose behavior determines the money stock and its close substitutes, as well as determining the interest rates in the economy. The institutions supplying the main components of the money stock are the central bank and the commercial banks. The commercial banks are themselves part of the wider system of financial intermediaries, which determine the supply of some of the components of money as well as the substitutes for money, also known as near-monies.

The two major components of the microeconomics part of monetary economics are the demand for money, covered in Chapters 4 to 9, and the supply of money, covered in Chapter 10. The central bank and its formulation of monetary policy are covered in Chapters 11 and 12.

Macroeconomics part of monetary economics: money in the macroeconomy

The macroeconomics part of monetary economics is closely integrated into the standard short-run macroeconomic theory. The reason for such closeness is that monetary phenomena are pervasive in their influence on virtually all the major macroeconomic variables in the short-run. Among variables influenced by the shifts in the supply and demand for money are national output and employment, the rate of unemployment, exports and imports, exchange rates and the balance of payments. And among the most important questions in macroeconomic analysis are whether – to what extent and how – the changes in the money supply, prices and inflation, and interest rates affect the above variables, especially national output and employment. This part of monetary economics is presented in Chapters 13 to 20.

A departure from the traditional treatment of money in economic analysis is provided by the overlapping generations models of money. These have different implications for monetary policy and its impact on the economy than the standard short-run macroeconomic models.

The long-run analysis of monetary economics is less extensive and, while macroeconomic growth theory is sometimes extended to include money, the resulting monetary growth theory is only a small element of monetary economics.

There are different approaches to the macroeconomics of monetary policy. These include the models of the classical paradigm (which encompass the Walrasian model, the classical and neoclassical models) and those of the Keynes's paradigm (which encompass Keynes's ideas, the Keynesian models and the new Keynesian models). We elucidate their differences at an introductory level towards the end of this chapter. What is money and what does it do?

MONETARY THEORY AND POLICY

The monetary aspects of the traditional classical approach were encapsulated in the quantity theory for the determination of the price level and the loanable funds theory for the determination of the interest rate. The statement of the quantity theory was an evolutionary one, with several – at least three – quite distinct approaches to the role of money in the economy. These quite diverse approaches shared the common conclusion that, in equilibrium, changes in the money supply caused proportionate changes in the price level but did not change output and unemployment in the economy. One of these approaches, provided by Knut Wicksell, proved to be a precursor of several aspects of the Keynesian macroeconomic approach.

The Keynesian approach discarded the quantity theory and integrated the analysis of the monetary sector and the price level into the complete macroeconomic model for the economy. For the monetary sector, it elaborated on the motives for holding money, leading to the modern approach to the analysis of the demand for money.

Key concepts introduced in this chapter

- ♦ An identity versus a theory
- ♦ Quantity equation
- ♦ Quantity theory
- ♦ Wicksell’s pure credit economy
- ♦ Transactions demand for money
- ♦ Speculative demand for money
- ♦ Precautionary demand for money
- ♦ Transmission mechanism
- ♦ Direct transmission mechanism
- ♦ Indirect transmission mechanism
- ♦ Lending channel
- ♦ Permanent income

The discussion of the role of money in the determination of prices and nominal national income in the economy has chronologically an extremely long heritage, extending back to Aristotle in ancient Greece, with explicit formulation of theories on it emerging in the mid-17th century. Current monetary theory has evolved from two different streams: the quantity theory stream, which was a part of the classical set of ideas, and the Keynesian one. This heritage includes both the microeconomic and macroeconomic aspects of monetary economics.

The quantity theory is the name given to the ideas on the relationship between the money supply and the price level from the middle of the eighteenth century to the publication of Keynes's *The General Theory* in 1936. It was a fundamental part of the traditional classical approach in economics. The specification of the quantity theory was an evolutionary tradition with several – at least three – distinct approaches to the role of money in the economy. These quite diverse approaches shared the common conclusion that, in long-run equilibrium, the changes in the money supply caused proportionate changes in the price level but did not change output or unemployment in the economy. The three approaches to the quantity theory are those based on the quantity equation (see Fisher's [1911] version of this approach below), on the demand for money in the Cambridge (UK) tradition (see Pigou's [1917] version of this approach below) and on a broader macroeconomic analysis (see Wicksell's [1907] approach below). Of these, the demand-for-money approach led to Keynes's elaboration of money demand, and Wicksell's approach led to both Keynes's and the current new Keynesian macroeconomic determination of the price level in a general macroeconomic framework.

The Keynesian approach discarded certain aspects of the quantity theory ideas and developed others in a new and distinctive format. On the demand for money, it elaborated on the earlier Cambridge approach and also rearranged its presentation in terms of the motives for holding money. This treatment in terms of motives eventually led to the modern treatment of the demand for money in terms of four motives: transactions, speculative, precautionary and buffer stock. The Keynesian emphasis on money as an asset, held as an alternative to bonds, also led to Friedman's analysis of the demand for money as an asset, thereby bringing this approach to money demand into the folds of the classical paradigm. At the macroeconomic level, Keynesian analysis made commodity market analysis, based on consumption, investment and the multiplier, a core part of macroeconomics. In doing so, it followed Wicksell. The Keynesian approach also integrated the analysis of the monetary sector into the complete macroeconomic model for the economy.

This chapter's very brief review of this heritage covers the contributions of David Hume, Irving Fisher, A.C. Pigou and Knut Wicksell for the classical period in economics and of John Maynard Keynes and Milton Friedman for the post-1936 period. In the evolution of ideas, the theoretical and empirical analysis of the demand for money only emerged during the twentieth century as a major element of monetary economics. This chapter reviews the three approaches to the quantity theory, followed by the contributions of Keynes and Friedman on the demand for money. It ends with the review of the transmission channels through which changes in the money supply affect aggregate demand and output.

Quantity equation

Any exchange of goods in the market between a buyer and a seller involves an expenditure that can be specified in two different ways.

A. Expenditures by a buyer must *always* equal the amount of money handed over to the sellers, and expenditures by the members of a group which includes both buyers and sellers must *always* equal the amount of money used by the group, multiplied by the number of times it has been used over and over again.[1] Designating the average number of times money turns over in financing transactions as its velocity of circulation V, expenditures as $\$Y$ and the money stock in use as $\$M$, we have $\$Y \equiv \$MV$, where $\equiv$ indicates an *identity* rather than merely an equilibrium condition.

B. Expenditures on the goods bought can also be measured as the quantity of physical goods traded times the average price of these commodities.[2] Expenditures Y then always equal the quantity y of the goods bought times their price level P, so that $\$Y \equiv \$Py$.

Obviously, these two different ways of measuring expenditures must yield the identical amount. These two measures are:

$$Y \equiv MV$$

$$Y \equiv Py$$

Hence,

$$MV \equiv Py \tag{1}$$

where:

y = real output (of commodities)
P = price level (i.e. the average price level of commodities)
Y = nominal value of output ($\equiv$ nominal income)
M = money supply
V = velocity of circulation of money (M) against output (y) over the designated period.

Equation (1) is an identity since it is derived solely from identities. It is valid under any set of circumstances whatever since it can be reduced to the statement: in a given period, by a given group of people, expenditures equal expenditures, with only a difference in the computational method between them. (l) is *true* for any person or group of persons.[3] If it is applied, as it usually is, to the aggregate level for the whole economy, the two sides of the identity and its four variables refer to all expenditures in the economy. But if it is applied to the world economy as a whole, its total expenditures and the four variables will be for the world economy.

(1) is called the *quantity equation*, the word "equation" in this expression serving to distinguish it from the *quantity theory,* which is vitally different in spirit and purpose from the quantity equation. As we shall see later, the quantity theory is not an identity, while the

quantity equation is not a *theory* for the determination of prices, incomes or even the velocity of circulation in the economy.

Note that a relationship or statement that is *always* valid under *any* circumstances is said to be an *identity* or *tautology*. Identities generally arise by the way the terms in the relationship are defined or measured. Thus, (1) defines (measured) expenditures in two different ways, once as *MV* and then as *Py*, so that (1) is an identity. An identity is different from an equilibrium condition that holds only if there is equilibrium but not otherwise – i.e. when there is disequilibrium. Further, a *theory* may or may not apply to any particular economy in the real world or it may be valid for some states – e.g. equilibrium ones – but not for others, while an *identity* is true (or false) by virtue of the definitions of its variables and its logic, so that its truth or falsity cannot be checked by reference to the real world. A theoryusually includes some identities but must also include behavioral conditions – which are statements about the behavior of the economy or its agents – and often also equilibrium conditions on its markets.

Note also that the velocity of circulation *V* depends on the length of the period of analysis. Since *Y* is a flow while *M* is a stock, the longer the period of analysis, the larger will be *Y* whereas *M* will be a constant. Therefore, *V* will increase with the length of the period.

Policy implications of the quantity equation for persistently high rates of inflation

Rewrite the quantity equation in terms of growth rates as:

$$M^{JJ} + V^{JJ} \equiv P^{JJ} + y^{JJ}$$

where ^{JJ} indicates the rate of change (also called the growth rate) of the variable. This identity can be restated as:

$$\pi \equiv M^{JJ} + V^{JJ} - Y^{JJ}$$

where π is the rate of inflation and is the same as P^{JJ}. This identity asserts that the rate of inflation is always equal to the rate of money growth plus the growth rate of velocity less the growth rate of output. *Ceteris paribus*, the higher the money growth rate, the higher will be the inflation rate, whereas the higher the output growth rate is, the lower will be the inflation rate. Note that velocity also changes over time and can contribute to inflation if it increases, or reduce inflation when it falls.[4]

In normal circumstances in the economy, velocity changes during a year but not by more than a few percentage points. Similarly, for most economies, real output growth rate is usually only a few percentage points. For the quantity equation, we need only consider the difference ($V^{JJ} - y^{JJ}$) between them. In the normal case, both velocity and output increase over time but the difference in their growth rates is likely to be quite small, usually in low single digits. *Adding this information to the quantity equation* implies that high (high single digits or higher numbers) and persistent (i.e. for several years) rates of inflation can only stem from high and persistent money growth rates. This is particularly true of hyperinflations in which the annual inflation rate may be in double (10 percent or more) or triple (100 percent or more) digits or

[1] The spread of banks and automatic teller/banking machines (ATMs) has tended to increase velocity in recent decades.

even higher. Empirically, even at low inflation rates, the correlation between money supply growth and inflation rates over long periods is close to unity.

To reiterate, the source of inflation over long periods is usually money supply growth and the source of persistently high inflation over even short periods is high and persistent money growth rates. Therefore, if the monetary authorities wish to drastically reduce inflation rates to low levels, they must pursue a policy that achieves an appropriate reduction in money supply growth.

Some variants of the quantity equation

There are several major variants of the quantity equation. One set of variants focuses attention on the goods traded or the transactions in which they are traded, so that they modify the right-hand side of (1). The second set of variants imposes disaggregation on the media of payments (e.g. into currency and demand deposits) or changes the monetary aggregate, thereby modifying the left side of (1). We present some forms of each of these variants. The first set of these variants is given by (i) and (ii) below. The second set is given by(iii).

(i) Commodities approach to the quantity equation

One way of measuring expenditures is as the multiple of the amount *y* of commodities sold in the economy in the current period times their average price level *P*. Therefore, the quantity equation can be written as:

$$M \cdot V_{My} \equiv P_y \cdot y \tag{2}$$

where:

V_{My} = income-velocity of circulation of money balances *M* in the financing of the commodities in *y* over the designated period

P_y = average price (price level) of currently produced commodities in the economy

y = real aggregate output/income in the economy.

(2) is often also stated as:

$$M \cdot V_{My} \equiv Y \tag{3}$$

(3) yields velocity V_{My} as equaling the ratio Y/M.

(ii) Transactions approach to the quantity equation

If the focus of the analysis is intended to be the number of *transactions* in the economy rather than on the quantity of goods, expenditures can be viewed as the number of transactions *T* of all goods, whether currently produced or not, in the economy times the average price P_T paid *per transaction*. The concept of velocity relevant here would be the rate of turnover per period of money balances in financing all such transactions. The quantity equation then becomes:

$$M \cdot V_{MT} \equiv P_T \cdot T \tag{4}$$

where:

V_{MT} = transactions-velocity of circulation per period of money balances M in financing transactions T

P_T = average price of transactions

T = number of transactions during the period.

To illustrate the differences between y and T and between P_y and P_T , assume that we are dealing with a single transaction involving the purchase of ten shirts at a price of $10 each. The total cost of the transaction is $100. Here, the quantity y of goods is 10 and their average price P_y is $10, while the number of transactions T is one and their average price P_T is $100.

(iii) Quantity equation in terms of the monetary base

The monetary base[5] consists of the currency in the hands of the public (households and firms), the currency held by the financial intermediaries and the deposits of the latter with the central bank. Since the central bank has better control over the monetary base, which it can manipulate through open market operations, than over M1 or M2, it is sometimes useful to focus on the velocity of circulation $V_{M\ 0,y}$ of the monetary base. This velocity depends not only upon the behavior of the non-banking public but also upon the behavior of firms and financial intermediaries. The quantity equation in terms of the monetary base is:

$$M0 \cdot V_{M0,y} \equiv P_y \cdot y \tag{5}$$

where:

$M0$ = quantity of the monetary base

$V_{M\ 0,y}$ = income-velocity of circulation per period of the monetary base.

The quantity equation is thus a versatile tool. Note that all versions of it are identities. The form in which it is stated should depend upon the analysis that is to be performed. Examples of such interaction between the intended use and the actual variant of the quantity equation employed occur often in monetary economics.

Quantity theory

The quantity theory had a rich and varied tradition, going as far back as the eighteenth century. It is the proposition that *in long-run equilibrium, a change in the money supply in the economy causes a proportionate change in the price level, though not necessarily in disequilibrium*.

The quantity theory was dominant in its field through the nineteenth century, though more as an approach than a rigorous theory, varying considerably among writers and periods. Two versions of the form that it had achieved by the beginning of the twentieth century are presented below from the works of Irving Fisher and A.C. Pigou. A third version, radically different from those of these writers, is presented later from the writings of Knut Wicksell.

Transactions approach to the quantity theory

Irving Fisher, in his book *The Purchasing Power of Money* (1911), sought to provide a rigorous basis for the quantity theory by approaching it from the quantity equation. He recognized the latter as an identity and added assumptions to it to transform it into a theory for the determination of prices. A considerable part of his argument was concerned with providing a clear and relevant exposition of the quantity equation, and one of his versions of this equation is presented below.

Fisher distinguished between currency and the public's demand deposits in banks. This distinction was relevant to the economy when he wrote, since currency was commonly used in payments whereas payments by check were much less common. Setting aside this distinction for the modern economy, we use M1 as the relevant money variable. Fisher also stated his version of the quantity theory in terms of the number of transactions, rather than in terms of the quantity of commodities purchased.[6] However, as a result of Keynes's emphasis on national income/output rather than total transactions, while data on national income/output came to be gathered and made commonly available, the data on the number of transactions was not gathered and has not become available in the public domain. The following, therefore, adapts Fisher's treatment of the quantity equation and theory and couches it in terms of the amount of the commodities purchased rather than in terms of transactions. This adapted version of the quantity equation has the form:

$$MV \equiv Py \tag{6}$$

To transform the quantity equation into the quantity theory, Fisher put forth two propositions about economic behavior. These are:

> (i) The velocities of circulation of "money" (currency) and deposits depend … on technical conditions and bear no discoverable relation to the quantity of money in circulation. Velocity of circulation is the average rate of "turnover" and depends on countless individual rates of turnover. These … depend on individual habits. … The average rate of turnover … will depend on density of population, commercial customs, rapidity of transport, and other technical conditions, but *not on the quantity of money and deposits nor on the price level*.
>
> (ii) *(except during transition periods)* the volume of trade, like the velocity of circulation of money, is *independent of the quantity of money*. An inflation of the currency cannot increase the product of farms and factories, nor the speed of freight trains or ships. The stream of business depends on natural resources and technical conditions, not on the quantity of money. The whole machinery of production, transportation and sale is a matter of physical capacities and technique, none of which depend on the quantity of money.
>
> (Fisher, 1911).

Therefore, Fisher's conclusion was that:

> while the equation of exchange is, if we choose, a mere "*truism*" based on the equivalence, in all purchases, of the money … expended, on the one hand, and what they buy on the other, yet *in view of supplementary knowledge … as to the non-relation of* [*velocity to money and prices*], this equation is the means of demonstrating the fact that normally the *prices vary directly as M*, that is, demonstrating the quantity theory.
>
> (Fisher, 1911, italics and the clause in brackets added).[7]

Fisher was certainly right in specifying that the transformation from his version of the quantity equation to the quantity theory *requires* that, when the monetary authorities increase the amount of money, the velocity of circulation and the quantities of goods remain unchanged. These assertions, as well as (i) and (ii) above, are economic ones, resting on assumptions about human behavior, and may or may not be valid. In symbols and in the above updated mode of statement of the quantity equation, these assertions become: $\partial y/\partial \underline{M}$ 0 and $\partial V/\partial M$ $\underline{0}$. These imply that, following an increase in the money supply, prices will rise in proportion to the increase in the money supply. That is, the elasticity of the price level with respect to the money supply will be unity.[8]

Fisher pointed out that the above assertions did not necessarily apply during "transition" (which can be interpreted as "disequilibrium") periods, so that his assertions applied to a comparison of the equilibrium states prior to and after a one-time increase in the money supply. Fisher based these assertions on the then dominant theories of output and other real variables (including velocity), for which the traditional classical approach and Walrasian model imply the independence of real variables from the monetary ones, which are *M* and *P*, in equilibrium.

On assumption (ii) of Fisher, the dominant theory – which was part of Fisher's own views of the economy – of the early twentieth century on output and employment in the economy was the Walrasian one, which treated each market separately and used microeconomic analysis.[9] This analysis implied that the labor market would clear in equilibrium and there would be full employment. Output would tend to stay at the full-employment level, except in the transient disequilibrium stages. Further, this full-employment output was independent of the money supply and prices. Therefore, Fisher's assertion that changes in the money supply would not affect the equilibrium output of goods was consistent with the real economic theory of the time and was, in effect, based on the latter. This assertion was to be later challenged by Keynes and the Keynesians for demand-deficient economies, reaffirmed by the modern classical economists in the 1980s and 1990s and denied by the new Keynesians in the last two decades.[10] Further, note Fisher's qualification "*except during transition periods*" to the quantity theory proposition. Interpreting this as a reference to the disequilibrium induced by an exogenous change in the money supply, the real-time of this transition (from one long-run equilibrium state to the one following a change in the money supply) becomes a very relevant question for the pursuit, or not, of monetary policy.

Fisher's assumption (i) on the independence of velocities from changes in the money supply is also questionable. The velocity of circulation of money is not directly related to the behavior of firms and households and, if one thinks solely in terms of velocity, Fisher's simplistic argument on this point seems reasonable. However, since velocity is a ratio of expenditures to money holdings, Fisher's assertion becomes more easily subject to doubt if the determinants of velocity are approached from the determinants of expenditures and the demand for money, as Keynesians do, and if the economy is not continuously in general equilibrium at full employment. These determinants include interest rates and output, so that changes in interest rates and in output can change both the demand for money and its velocity.

However, velocity is a real variable since it can be defined as equal to real income divided by the real money stock in the economy. Modern classical economists focus on velocity as a real variable, as Fisher had done, and, along with other real variables, take it to be independent of money supply and the price level in the long-run equilibrium state of the economy. Hence, modern classical economists agree with both of Fisher's assumptions for the general equilibrium – that is, with all markets clearing – state of the economy. Modern classical economists, therefore*, with a model implying continuous full employment*, still maintain Fisher's quantity theory assertion that an increase in the money supply will cause a proportionate increase in the price level, with velocity remaining unchanged.

Keynesians question the empirical usefulness of the assumption of continuous long-run general equilibrium (yielding full employment) since they maintain that continuous full employment does not normally exist in the economy. They also assert the dependence of money demand on the interest rate and the dependence of the interest rate on liquidity preference and the money supply. Hence, to the Keynesians, neither velocity nor output is independent of the money supply. Therefore, Keynesians reject the validity of the quantity theory both in terms of comparison across equilibrium states and in disequilibrium.

Determinants of velocity: constancy versus the stability of the velocity function

Equilibrium in the money market means that money demand equals money supply. Therefore, in this equilibrium, velocity can be redefined as the nominal income divided by money demand. As explained in subsequent sections of this chapter, money demand depends upon many variables, of which the most important are national income and interest rates. As income rises, economies of scale in money holdings mean that money demand does not rise as fast, so that velocity increases. The interest rate is the cost of holding money rather than interest-paying financial assets, so that money demand falls as interest rates rise, which increases velocity. Therefore, velocity rises as income rises and also rises as interest rates rise.

Financial innovations in recent decades have created a variety of substitutes for M1 and M2, which have reduced their demand. Further, telephone and electronic banking have reduced the need to hold large precautionary balances against unexpected needs for expenditures. This trend has been reinforced by the fall in brokerages costs of various types in switching between money and other financial assets, so that individuals can manage their expenditures with smaller money balances while holding larger amounts of interest-paying financial assets. These developments have reduced the demand for M1 and M2, so their velocity has risen considerably in recent decades. To illustrate, while the velocity of M1 in the USA was about 6.3 in 1991, it rose to about 8.8 in 2000.

Table 2.1 shows that the velocity of circulation, which equals nominal national income divided by the money supply, in Canada, UK and USA is not a constant. In fact, it varies even over periods as short as a day or month.

Fisher did not assume the constancy of velocity. His assumption was the independence of velocity – a real variable – from that of changes in the money supply and the price level in the general equilibrium states of the economy. From an empirical perspective, velocity is not a constant in either the short term or the long term in actual economies. It is continuously changing in the economy. Some estimates of the average annual change in velocity for the USA lie at about 3 percent to 4 percent.

To conclude, Fisher did not assume velocity to be a constant, nor is it constant in the real economy. Economic theory takes it to be an economic variable, determined in the economy by other economic variables. As its determinants change, velocity changes. The determinants of velocity are discussed in greater detail later in this chapter.

The Fisher equation on interest rates: distinction between nominal and real interest rates

Another of Fisher's contributions on monetary theory was his distinction between the nominal and real interest rates. This is embodied in what has been designated the Fisher equation.

The rate of interest that is charged on loans in the market is the *market* or *nominal rate of interest*. This has been designated by the symbol R. If the rational lender expects a rate of inflation π^e, he has to consider the real interest rate r that he would receive on his loan. However, financial markets usually determine the nominal interest rate R. In perfect capital markets, the *ex ante* relationship[14] between the expected real interest rate r^e and the nominal interest rate R is specified by

$$(1 + r^e) = (1 + R)/(1 + \pi^e) \tag{7}$$

where π^e is the expected inflation rate. If there exist both real bonds (i.e. promising a real rate of return r per period) and nominal bonds (i.e. promising a nominal rate of return Rper period), the relationship between them in perfect markets would be:

$$(1 + R) = (1 + r)(1 + \pi^e) \tag{8}$$

At low values of r^e and π^e, $r^e\pi^e \to 0$, so that (7) is often simplified to:

$$r^e = R - \pi^e \tag{9}$$

This states that the real yield that the investor expects to receive equals the nominal rate minus the expected loss of the purchasing power of money balances through inflation. (8) is correspondingly simplified to $R\ r\ \pi^e$. (8) and (9) are known as the Fisher equation.

Note that the real value of the rate of return that the holder of a nominal bond would actually (i.e. *ex post*) receive from his loan is the *actual real rate of interest* (r^a), which is correspondingly given by:

$$r^a = R - \pi \tag{9^J}$$

In these equations, the definitions of the symbols are:

R = nominal rate of interest
r^a = actual (*ex post*) real rate of interest on nominal bonds
r = real rate of interest
r^e = expected real rate of return
π = actual rate of inflation
π^e = expected rate of inflation.

[2] One explanation for the Fisher equation is as follows. An investor investing one dollar in a "nominal bond" (i.e. paying a nominal rate of interest R) would receive \$$(1 + R)$ at the end of the period. If he were to buy a "real bond" (i.e. paying a real interest rate r), he would receive $(1 + r)$ in real terms (i.e. in commodities) at the end of the period. Given the expectations on inflation held at the beginning of the current period, the expected nominal value at the end of the period of this real amount equals \$$(1 + r)(1 + \pi^e)$. The investor would be indifferent between the nominal and the real bonds if the nominal return from both bonds were equal, i.e. $(1 + R) = (1 + r)(1 + \pi^e)$. With all investors behaving in this manner, perfect capital markets would ensure this relationship.

If the actual rate of inflation were imperfectly anticipated, the actual yield r^a on nominal bonds would differ from the expected one r^e and may or may not be positive. In fact, negative real interest rates are often observed during years of accelerating inflation, such as in the 1970s, when the real yield on nominal bonds was often, and often persistently, negative.[15]

Fisher's direct transmission mechanism

For the transmission mechanism from exogenous money supply changes to the endogenous changes in aggregate demand and prices, Fisher argued that an increase in the money supply leads its holders to increase their expenditures on commodities. Fisher's version of this disequilibrium chain of causation from changes in the money supply to changes in the nominal value of aggregate expenditures is given in the following quotation. Fisher starts by assuming that an individual's money holdings are doubled, and continues as:

> Prices being unchanged, he now has double the amount of money and deposits, which his convenience had taught him to keep on hand. *He will then try to get rid of the surplus money and deposits by buying goods*. But as somebody else must be found to take the money off his hands, its mere transfer will not diminish the amount in the community. It will simply increase somebody else's surplus.… Everybody will want to exchange this relatively useless extra money for goods, and the desire so to do must surely drive up the price of goods. [This process will continue until prices double and equilibrium is restored at the initial levels of output and velocity.]
>
> (Fisher, 1911, italics added).

Fisher's mechanism, by which changes in the money supply induce changes in aggregate expenditures, has come to be known as the *direct transmission mechanism* of monetary policy, as compared with the *indirect transmission mechanism*, which relies upon the changes in the money supply inducing changes in interest rates, which in turn induce changes in investment, which then cause changes in aggregate expenditures. The latter mechanism was incorporated in the 1930s into the Keynesian and neoclassical macroeconomic models, but the former was revived by Milton Friedman and the 1970s monetarist models. The modern classical models generally ignore the direct transmission mechanism and, as with the Keynesian models, incorporate the indirect transmission mechanism. However, the direct transmission mechanism continues to be relevant to the poor whose expenditures are close to their incomes, and especially in economies in which the increase in the money supply is used to finance fiscal deficits and initially ends up in the hands of people whose usual use of extra funds is to buy commodities.

Cash balances (Cambridge) approach to the quantity theory

Another popular approach to the quantity theory examined the determination of prices from the perspective of the demand and supply of money. Some of the best known exponents of this approach were at Cambridge University in England and included, among others, Alfred Marshall, A.C. Pigou and the early writings (that is, pre-1936) of John Maynard Keynes. The following exposition of this approach follows that of Pigou in his article, *The Value of Money* (1917).

Pigou, like Fisher, defined currency or *legal tender* as money but was, in general, concerned with what he called "*the titles to legal tender.*" He defined these titles as including currency and demand deposits in banks, which correspond to the modern concept of M1. He argued that a person held currency and demand deposits:

> to enable him to effect the ordinary transactions of life without trouble, and to secure him against unexpected demands due to a sudden need, or to a rise in the price of something that he cannot easily dispense with. For these two *objects*, the *provision of convenience* and the *provision of security*, people in general elect to hold currency and demand deposits.
>
> (Pigou, 1917, italics added).

The actual demand for currency and demand deposits is:

> determined by the *proportion* of his resources that the average man chooses to keep in that form. This proportion depends upon the convenience obtained and the risk avoided through the possession of such titles, by the loss of real income involved through the provision to this use of resources that might have been devoted to the production of future commodities, and by the satisfaction that might be obtained by consuming resources immediately and not investing at all.
>
> (Pigou, 1917).

Pigou thus claimed that the individual is not directly concerned with the demand for money but with its relation to his total resources. These resources can be interpreted as wealth in stock terms or as income/expenditures in flow terms. We will use the latter, so that income will be the proxy for Pigou's "resources." Further, according to Pigou, this ratio of money demand to resources is a function of its services, the internal rate of return on investments and of the marginal satisfaction foregone from less consumption. Representing the internal rate of (real) return on investment as r and assuming it to be an approximate measure, in equilibrium, of the satisfaction foregone by not consuming, the ratio of money balances demanded (M^d) to total nominal expenditures (Y) is given by:

$$M^d/Y = k(r) \quad k'(r) < 0 \tag{10}$$

where k is a *functional* symbol. M^d/Y decreases with r, or, in Pigou's words, "the variable k will be larger the less attractive is the production use and the more attractive is the rival money use of resources." Hence, $\partial k/\partial r < 0$. Therefore, the demand for money balances, M^d, is:

$$M^d = k(r)Y \tag{11}$$

Determination of the price level in the cash balance approach

Assuming a given money supply M, equilibrium in the money market with (11) requires that:

$$M = k(r)Y \tag{12}$$

Writing Py for Y, with P as the price level and y as the real amount of goods,

$$M = k(r)\,Py \tag{13}$$

Assuming that output y is at its full employment level y^f in equilibrium, $y = y^f$, so that (11) becomes:

$$M = k(r)\,Py^f$$

where $\partial y^f/\partial P = 0$ and $\partial y^f/\partial M = 0$. Further, Pigou assumed[16] that the equilibrium rate of return (r^*) was determined by the marginal productivity of capital (MPK), which was taken to be independent of the money supply and the price level, so that $\partial r^*/\partial P = 0$ and $\partial r^*/\partial M = 0$. Therefore, in equilibrium,

$$M = k(r^*)\,Py^f \tag{14}$$

so that, in equilibrium,

$$P = M \big/ \left[k\left(r^*\right) \cdot y^f \right] \tag{15}$$

which implies that:

$$\partial P/\partial M = 1 \big/ \left[k\left(r^*\right) \cdot y^f \right]$$

and

$$E_{P\cdot M} = (M/P) \cdot (\partial P/\partial M) = 1$$

where $E_{P\,M\,.}$ is the elasticity of P with respect to M. Since this elasticity equals unity, the price level will, in *comparative static equilibria*, vary proportionately with the money supply. Therefore, (14) establishes Pigou's version of the quantity theory proposition.

The cash balance approach starts its statement of the quantity theory as a theory of demand, supply and equilibrium in the money market and then proceeds to place it in a long-run general equilibrium approach to the economy. From a rigorous standpoint, it does not become a theory of the price level until the complete model – which includes the determination of output and interest rates – is specified. On the latter variables, Pigou and his colleagues in the quantity theory tradition had in mind the then generally accepted traditional classical ideas on the determination of output and interest rates. As stated in Chapter 1, these ideas implied the

[3] Pigou implicitly did so in *The Value of Money*. This was consistent with his ideas on the determination of the equilibrium rate of return in the economy by the marginal productivity of capital.

independence of the long-run equilibrium values of both these variables from the demand and supply for money and turned the money market equilibrium equation (11) into a statement of the quantity theory. The essential deficiency in Pigou and the cash balance approach lay not so much in their specification of money demand relevant to the time in which they were writing, but in that of the existing (traditional classical) macroeconomic analysis which failed to specify the determination of aggregate demand and its impact on output in short-run equilibrium, as well as in disequilibrium. This deficiency was the major point of attack by Keynes on the quantity theory and the traditional classical approach generally.

Velocity in the cash balance approach

On the velocity of circulation V in Pigou's analysis, we have from (11) that:

$$\begin{aligned} V &= Y/M \\ &= 1/[k(r)] \qquad (16) \end{aligned}$$

In (13), since velocity depends upon the rate of interest, it is not a constant in the context of Pigou's money market analysis. However, given the independence of the equilibrium rate of interest and the marginal productivity of capital from the supply of money, the *equilibrium* level of velocity equals $[1/k(r^*)]$, which is independent of the supply of money. This independence of velocity with respect to the money supply does not mean its constancy over time, since velocity could still depend upon other variables, such as banking practices and payment habits, and these often change over time. Further, the independence of velocity from the money supply was asserted only for equilibrium but not for disequilibrium. However, Pigou and other economists in the Cambridge school often fell into the habit of treating k as a constant even though it was a functional symbol with $k^J(r) < 0$, so that velocity also became a constant both in and out of equilibrium.

Legacy of the cash balance approach for the analysis of the demand for money

Further developments in monetary theory during the twentieth century built on two aspects of the nineteenth and early twentieth century monetary theory. These were as follows: (i) The cash balance approach started its presentation of the quantity theory by analyzing the demand for money and equilibrium in the money market. This idea was later taken up by Milton Friedman (whose contribution on this topic is presented later in this chapter) to identify and confine the quantity theory to the analyses of the demand for money and the money market; (ii) the cash balance approach had analyzed the demand for money in terms of its characteristics or functions, which were:

1 *The provision of convenience in transactions*.

2 *The provision of security against unexpected demands due to a sudden need or to a rise in the prices.*

The former was related to the demand for the medium of exchange function of money and the latter to its store of value function. These reasons for holding money were restated by Keynes in 1936 into the transactions motive and the precautionary motive. Keynes added to these the speculative motive.

Wicksell's pure credit economy

Knut Wicksell was a Swedish monetary economist writing within the classical tradition in the last decades of the nineteenth and the first quarter of the twentieth century and considered himself to be an exponent of the quantity theory. His treatment of the quantity theory was very distinctive and quite different from the English and American traditions of the time, as represented in the works of Fisher, Pigou and Keynes during his classical period prior to 1930. Further, elements of Wicksell's analysis led to the formulation of modern macroeconomic analysis. His ideas have assumed even greater importance in the past two decades since several central banks in developed economies have adopted the use of the interest rate as their primary monetary policy instrument, so that the appropriate analysis has to take the interest rate rather than the money supply as being exogenously set. The money supply becomes endogenous in this context. These assumptions are essentially similar to those made by Wicksell. The new Keynesian analysis embodies these assumptions, so that it is sometimes referred to as the neoWicksellian analysis.

Wicksell sought to defend the quantity theory as the appropriate theory for the determination of prices against its alternative, the *full cost pricing* theory. The latter argued that each firm sets the prices of its products on the basis of its cost of production, including a margin for profit, with the aggregate price level being merely the average of the individual prices set by firms. The amount of the money supply in the economy adjusts to accommodate this price level and is therefore determined by the price level, rather than determining it. Wicksell considered this full cost pricing theory as erroneous and argued that such pricing by firms determined the relative prices of commodities, rather than the price level. In his analysis, the latter was determined by the quantity of money in the economy relative to national output since commodities exchange against money and not against each other.

In his reformulation of the quantity theory, Wicksell (1907) sought to shift the focus of attention to the transmission mechanism relating changes in the money supply to changes in the price level. He specified this mechanism for economies using either metallic or fiat money and for a *pure credit* economy. The latter analysis is the more distinctive one and illustrates Wicksell's transmission mechanism more clearly. It is also the one likely to be more relevant to the future evolution of our present day economies and, therefore, is the one presented below.

In modern macroeconomic terminology, Wicksell's analysis of the pure credit economy is essentially short run since his analysis assumes a fixed capital stock, technology and labor force in the production of commodities. This focus on the short run contrasts with Fisher's and Pigou's reliance on the long-run determination of output in order to establish their versions of the quantity theory. Further, Wicksell assumes that the economy is a pure credit one in the sense that the public does not hold currency and all transactions are paid by checks drawn on checking accounts in banks, which do not hold any reserves against their demand deposits. Since the banks do not hold reserves and any loans made by them are re-deposited by the borrowers or their payees in the banks, the banks can lend any amount that they desire without risking insolvency. Further, banks are assumed to be willing to lend the amount that the firms wish to borrow at the specified market rate of interest set by the banks. Wicksell calls the nominal rate of interest at which the banks lend to the public the *money* or *market rate of interest*. The banks accommodate the demand for loans at this interest rate, which is set by them. Under these assumptions, the amount of money supply in the economy is precisely equal to the amount of credit extended by the banks, since these loans are wholly deposited in the banks. Hence, changes in the money supply occur only when the demand for loans changes in response to an exogenous shift in the interest rate charged by banks. Note that, in Wicksell's pure credit economy, the economy's interest rate is set exogenously by the banks, while the money supply depends on this interest rate and the public's demand for loans. Therefore, it is endogenous to the economy.

A critical element of Wicksell's (1907) theory is the emphasis on saving and investment in the economy. Funds for (new) investment come from saving plus changes in the amount of credit provided by banks. The rate of interest which equates saving and investment was labeled by Wicksell the *normal rate of interest.* Since Wicksell's pure credit economy was a closed one and there was no government sector, the equality of saving and investment means that the normal rate of interest is the macroeconomic equilibrium rate. Further, if the market interest rate equals the normal rate, there will be no change in the credit extended by banks and, therefore, no change in the money supply. For a stable amount of credit and money supply in the economy, the price level will remain unaltered. To conclude, at the market rate of interest equal to the normal one, there is equilibrium in the commodity market, Further, with stable output and money supply, the normal rate of interest will be accompanied by a stable price level.

Firms borrow to finance additions to their physical capital. The marginal productivity of capital specifies the internal rate of return to the firm's investments and was referred to by Wicksell as the *natural rate of interest*. The firm's production function has diminishing marginal productivity of capital, so that, with a constant labor force and unchanged technology, the natural rate of interest decreases as capital increases in the economy.

To see the mechanics of this model, start from an initial position of equilibrium in the economy, with a stable money supply and prices, and with the equality of the market/loan and natural rates of interest at the normal/equilibrium rate of interest. Now, suppose that while the market rate of interest is held constant by the banks, the marginal productivity of capital rises. This could occur because of technological change, discovery of new mines, a fall in the real wage rate, etc. Firms can now increase their profits by increasing their capital stock and production. To do so, they increase their investments in physical capital and finance these by increased borrowing from the banks. This causes the amount of credit and money supply in the economy to expand.

Wicksell appended to this analysis the disaggregation of production in the economy between the capital goods industries and the consumer goods industries. As the demand for investment in physical capital increases, factors of production are drawn into such industries from the consumer goods industries, so that the output of the latter falls. At the same time, the competition for labor and the other factors of production will drive up workers' incomes, leading to an increase in the demand for consumer goods, thereby pushing up prices. Consequently, the price level will rise, though with a lag behind money supply changes. Analysis based on this disaggregation of production between the capital goods industries and the consumer goods industries is not a feature of most modern macroeconomic models.

Cumulative price increases (the inflationary process)

In the above process, initiated by a reduction by the banks of the market interest rate below the natural one or by an increase in the latter above the market rate, the price rise will continue as long as the market rate of interest is below the natural rate, since the firms will then continue to finance further increases in investment through increased borrowing from the banks. This constitutes a process of cumulative price increases. These increases can only come to an end once the banks put an end to further increases in their loans or credit to firms. A closed pure credit economy does not provide a mechanism that will compel the banks to do this.

However, in an open economy where the banking system keeps gold reserves out of which deficits in the balance of payments have to be settled, gold outflows provide a limit to the cumulative price increases. In such a context, as prices continue to increase, foreign trade deficits develop, the gold reserves of banks fall and the banks raise their loan rate of interest to the natural rate to stem the outflow of gold. This is especially so if the banks hold gold as part of their reserves and the public holds gold coins circulating as currency for some transactions. In the latter case, as prices rise, the public's demand for currency will also increase and gold will flow out of the banks' reserves to the public. Such losses of the gold reserves to the public and abroad forces banks to restrict their lending to the firms by raising their loan rate to match the natural rate. This puts an end to the cumulative credit and money supply increases and therefore to the cumulative price increases.

This cumulative process can also be initiated by banks arbitrarily lowering the market rate below the natural rate, with the resultant adjustments being similar to those specified above for an exogenous increase in the natural rate. However, Wicksell viewed the bankers as being conservative enough not to change the market rate except in response to changes in their gold holdings or an exogenous change in the normal rate. Therefore, in Wicksell's view, the cumulative price increase was usually a result of exogenous changes in the marginal productivity of capital impinging on an economy whose credit structure responds with gradual and possibly oscillatory adjustments – for example, if the banks sometimes overdo the adjustment of the market rate.

Wicksell's re-orientation of the quantity theory to modern macroeconomics

Wicksell's treatment of the pure credit economy clearly re-oriented the quantity theory in the direction of modern macroeconomic analysis. Several features of this analysis are relevant to modern macro and monetary economics. Among these is Wicksell's focus on the short-run treatment of the commodity market in terms of the equilibrium between saving and investment, a focus that was later followed and intensified in the Keynesian approach, as well as in the IS–LM modeling of short-run macroeconomics. While Wicksell claimed to be a proponent of the quantity theory of money, he shifted its focus away from exclusive attention on the monetary sector, for example, as in Pigou's version of the quantity theory, to the saving-investment process. In doing so, he led the way to the formulation of current macroeconomics, with the treatment of the commodities market at its core. This was to appear later as the IS relationship of modern macroeconomics.

Wicksell introduced into macroeconomics a fundamental aspect of the modern monetary economies: loans are made in money, not in physical capital, so that the rate of interest on loans is conceptually different from the productivity of physical capital. Even if they are equal in equilibrium, they will usually not be equal in disequilibrium. These ideas led the way to the analysis of the impact that the financial institutions and especially the central bank can have on the interest rates in the economy and on national income and employment.

Wicksell's analysis of the pure credit economy also emphasized the role of interest rates and financial institutions in the propagation of economic disturbances, since they control the market interest rate, reduction in which can set off an expansion of investment, loans and the money supply and lead to a cumulative increase in prices and nominal national income.

Further, Wicksell assumed that the banking system sets the interest rate rather than the money supply as the exogenous monetary constraint on economy. This assumption was not followed by the expositions of macroeconomic theory in either the classical or the Keynesian formulations until the end of the twentieth century, since they continued to take the money supply as their exogenously determined monetary policy variable. Since the money-demand function proved to be unstable in most developed economies after the 1970s, thereby implying the instability of the LM curve, many central banks now choose to use the interest rate as the monetary policy variable and set its level, while allowing the economy to determine the money supply as an endogenous variable for the set interest rate. This practice came to be reflected in the theories offered by the new Keynesian approach after the early 1990s. Wicksell was clearly the precursor of this type of analysis.

However, compared with the Keynesians, Wicksell, just like Fisher and Pigou, did not pay particular attention to the changes in the national output that might occur in the cumulative process. While he discussed disequilibrium and transient changes in national output during this process, he was not able to shake off the classical notion that the economy will eventually be at full employment, so that his overall discussion was usually within the context of an implicitly unchanged equilibrium level of output. Given this background, Wicksell claimed that increases in the money supply are accompanied sooner or later by proportionate price increases. Keynes's *General Theory* (1936) was to question the implicit assumption of an unchanged level of output and to allow for changes in output and unemployment following a change in aggregate demand. Merging this possibility into Wicksell's cumulative process would mean that his cumulative process would possess both output and price increases (decreases) whenever the market interest rate was below (above) the natural rate.

Hence, while Wicksell claimed nominal adherence to the traditional classical approach and the quantity theory, his theoretical macroeconomic analysis differed from theirs and was quite modern in several respects. One, in terms of this theoretical analysis in terms of saving and investment, Wicksell was a precursor of the Keynesian and modern short-run macroeconomic analysis. Two, in terms of his assumption of a pure credit economy, he presaged current developments in the payments system. Three, his assumption that the financial system sets the interest rate rather than the money supply as exogenous, he was a precursor of current central bank practices and the analysis of the new Keynesian models in the last couple of decades.

However, Wicksell's analysis did have at least several deficiencies relative to current monetary economics. One, although Wicksell did approach equilibrium through the normal interest rate which equates saving and investment, he did not present a theory of aggregate demand and also did not present the analysis of the impact of changes in it on output and employment. These were to be later addressed by Keynes. Two, Wicksell did not distinguish between real and nominal interest rates, which Fisher's equation later clarified. Three, he did not pay much attention to the analysis of the demand for money, on which Keynes made very significant contributions which provide the basis for its modern mode of treatment.

Keynes's contributions

Keynes's contributions to macroeconomics

Keynes's *The General Theory* (1936) represents a milestone in the development of macroeconomics and monetary thought. His contributions were so many and so substantial that they led to the development of the new field of macroeconomics, which had not existed in economic thought prior to *The General Theory*. These contributions also led to a new way of looking at the performance of the economy and to an emphasis on departures from its long-run equilibrium (full employment) and the establishment of the Keynesian paradigm (see Chapter 15) in macroeconomics.

Given the very many new contributions in this book, economists have debated as to which was the most important of these contributions.[17] From a modern perspective, Keynes's emphasis on aggregate demand as a major short-run determinant of aggregate output and employment seems to have had a lasting impact on economic theory and policy. Every presentation of macroeconomic theory now includes the determination of aggregate demand and its relationship, embodied in the IS curve, to investment and fiscal policy. This contribution was based on the concept of the multiplier, which was unknown in the traditional classical period. Keynes's impact on monetary policy is reflected in central banks' manipulation of aggregate demand through either the use of the money supply or/and the interest rate, in order to maintain inflation and output at their desired levels.

Again, in terms of the modern perspective, Keynes's emphasis was on decisions on production and investment being made by firms on the basis of their expectations of future demand, and on consumption by households on the basis of their expected incomes. These decisions are usually made under uncertainty, with imperfect information on the future. Following any shifts, the reactions by firms and households to changes in demand and income prospects are often faster than by heterogeneous commodity and labor markets in adjusting prices and wages, so that the economy often produces more or less than the long-run equilibrium (full employment) output that efficient (i.e. instantly adjusting) markets will ensure. The economy is, therefore, usually likely to end up with more or less than full employment. This provides the scope for the pursuit of monetary and fiscal economies to stabilize the economy. This scope is currently reflected in the espousal of Taylor-type rules for monetary policy.

Contrary to the assumptions of the quantity theory, *The General Theory* asserted the *usual* absence of full employment in the economy. This is clearly a factual issue, which is undeniable in the context of the Great Depression of the 1930s and in many recessions. In the context of actual employment below the full-employment level, Keynes argued that output and employment depended on the aggregate demand for commodities, which, in turn, depended on the money supply, so that money was not neutral. In the context of the lengthy post-war booms in the Western economies, the contribution of high and rising aggregate demand in pushing output and employment beyond their full-employment rates is also generally recognized. The current manifestation of this recognition can be seen in the pursuit by central banks of Taylor-type rules, in which the output gap can be positive (with output above its full-employment level) or negative, with appropriate increases and decreases in interest rates expected to reduce the output gap.

Keynes, in his earlier (pre-1936) writings, had proved to be an able and innovative exponent of the quantity theory in its Cambridge school version. He had also extensively explored the effects of changes in the money stock, though still mainly within the quantity theory tradition, in the two volumes of his book *The Treatise on Money,* published in 1930. Keynes's approach to the quantity theory in the *Treatise*, as in Wicksell's writings, was in terms of saving and investment. In *The General Theory,* Keynes extended this saving-investment

[17] Samuelson's (1946) obituary article on Keynes provides very valuable insights into Keynes's contributions.

approach, while abandoning the quantity theory and the traditional classical approach generally.

This chapter mainly examines Keynes's contributions on the demand for money in *The General Theory*. As a prelude to these, remember that Pigou's basic reasons for the demand for money balances were the "objects" of the provision of convenience and the provision of security. Keynes re-labeled "objects" as "motives" for holding money balances and categorized them as the transactions, precautionary and speculative motives. Of these, the transactions motive corresponded basically to the provision of convenience "object" of Pigou and the precautionary motive corresponded basically to the provision of security "object" of Pigou. Keynes was more original with respect to his speculative motive and his analysis of the demand for money balances arising from this motive.

Keynes's transactions demand for money

Keynes defined the transactions motive as:

> *The transactions-motive,* i.e. the need of cash for the current transaction of personal and business exchanges.
>
> (Keynes, 1936, Ch. 13, p. 170).

The transactions motive was further separated into an "income-motive" to bridge the interval between the receipt of income and its disbursement by households, and a "business-motive" to bridge the interval between payments by firms and their receipts from the sale of their products (Keynes, 1936, Ch. 15, pp. 195–6). Keynes did not present a rigorous analysis of the transactions and precautionary motives but "assumed [them] to absorb a quantity of cash which is not very sensitive to changes in the rate of interest as such … apart from its reactions on the level of income" (Keynes, 1936, p. 171). This assumption of Keynes was in fact somewhat more restrictive than that of Pigou where the demand for money, due to the objects of the "provision of convenience" and the "provision of security," was dependent upon the return on investments and the utility foregone in abstaining from consumption. Designating the *joint* transactions and precautionary demand for money balances as M^{tr} and nominal income as Y, Keynes assumed that:

$$M^{\mathrm{tr}} = M^{\mathrm{tr}}(Y) \tag{17}$$

where M^{tr} increases as Y increases.

Now consider the ratio (Y/M^{tr}), which is the velocity of circulation of transactions balances alone in the preceding equation. Here, Keynes followed the simplistic pattern of Pigou's reasoning in stating that

> There is, of course, no reason for supposing that V (= Y/M^{T}) is constant. Its value will depend on the character of banking and industrial organization, on social habits, on the distribution or income between different classes and on the effective cost of holding idle cash. Nevertheless, if we have a short period of time in view and can safely assume no material change in any of these factors, we can treat V as nearly enough constant.
>
> (Keynes, 1936, p. 201).

This reasoning implies that Y/M^{tr} is a constant k, independent of income and interest rates, so that Keynes's transactions demand for money was:

$$M^{\text{tr}} = kY \tag{18}$$

The modern analysis of transactions demand did not follow Keynes's simplistic assumption on its constancy, but applies inventory models to it, which makes this demand a function of the interest rate. This analysis is presented in Chapter 4.

Keynes's precautionary demand for money

Keynes's second motive for holding money was the precautionary one, defined by him as

> the desire for security as to the future cash equivalent of a certain proportion of total resources.
>
> (Keynes, 1936, Ch. 13, p. 170).

Another definition of this motive was given later in Chapter 15 of *The General Theory* as

> To provide for contingencies requiring sudden expenditure and for unforeseen opportunities of advantageous purchases, and also to hold an asset of which the value is fixed in terms of money.
>
> (Keynes, 1936, Ch. 15, p. 196).[18]

That is, the precautionary motive arises because of the uncertainty of future incomes, as well as of consumption needs and purchases. These require holding money, an asset with a certain value, to provide for contingencies that suddenly impose payment in money. These contingencies could come from a sudden loss of income due to the loss of one's job, or a sudden increase in consumption expenditures, such as from becoming ill and requiring treatment.

Under uncertainty, the individual will form subjective expectations on the amounts required for his future payments and income receipts, and their dates, and will decide on the optimal amounts of his money balances and other assets in the light of these expectations. The further ahead are the dates of anticipated expenditures and the greater is the yield from investments, the more likely is the individual to invest his temporarily spare funds in bonds and decrease his money holdings. Conversely, an increase in the probability of requirement in the near future will lead him to increase his money holdings and decrease his bond holdings.

Although Keynes provided the rationale for the precautionary motive for holding money, he did not present a theoretical derivation of the precautionary demand for money. Rather, he merged it with the transactions demand for money. However, subsequent developments on money demand did come up with several models of the precautionary demand for money and its related buffer stock demand (see Chapter 6).

Keynes's speculative money demand for an individual

Keynes's third motive for holding money was:

> 3. *The speculative-motive*, i.e. the object of securing profit from knowing better than the market what the future will bring forth.
>
> (Keynes, 1936, Ch. 13, p. 170).

Keynes had earlier explained this motive as resulting:

> from the existence of uncertainty as to the future of the rate of interest, provided that there is an organized market for dealing in debts. For different people will estimate the prospects differently and anyone who differs from the predominant opinion as expressed in market quotations may have a good reason for keeping liquid resources in order to profit, if he is right … the individual who believes that future rates of interest will be above the rates assumed by the market, has a reason for keeping liquid cash, whilst the individual who differs from the market in the other direction will have a motive for borrowing money for short periods in order to purchase debts of longer term. The market price will be fixed, at the point at which the sales of the "bears" and the purchases of the "bulls" are balanced.
>
> (Keynes, 1936, Ch. 13, pp. 169–70).

In this motive, the individual makes a choice between holding money, which does not pay interest, and bonds, which provide an uncertain return, on the basis of maximizing the return to his portfolio. With a given amount to invest in bonds or hold in money balances, he is concerned with the maturity value – equal to the capital invested plus accumulated interest – of his portfolio at the beginning of the next decision period. Assuming such a value to be uncertain, Keynes postulated a rather simple form of the expectations function: the individual anticipates a particular rate of interest to exist at the beginning of his next decision period, thereby implying a particular expected price, without dispersion,[19] for each type of bond. If these expected bond prices plus the accumulated interest are higher than the current prices, he expects a net gain from holding bonds, so that he will put all his funds in bonds rather than in money which was assumed not to pay interest and therefore to have zero net gain. If he expects a sufficiently lower price for bonds in the future than at present to yield a net loss[20] from holding bonds, he will put all his funds into money balances since there is no loss from holding these. Consequently, a particular *individual* will hold either bonds or money but not both simultaneously.

Since individuals tend to differ in their views on the future of the rate of interest, some would expect an increase in bond prices and are labeled as *bulls* in bond market parlance, choosing to increase their bond holdings, while others would expect a decrease in bond prices and are labeled as *bears*, choosing to reduce their bond holdings. Any increase in bond prices will exceed the expectations of some bulls – that is, convince them that bond prices have gone up too far and convert them into bears. A preponderance of bulls in the bond market pushes up the prices of the bonds and pushes down the rate of interest. This movement converts an

[18] This simplification was subsequently abandoned in the 1950s by monetary economics in the application of portfolio selection analysis to the speculative demand for money, presented in Chapter 5 below.

[19] There will be a net loss if the capital loss is greater than the interest income from holding the bond.

increasing number of bulls (who want to buy and hold bonds) into bears (who want to sell bonds and hold money), until an equilibrium price of bonds is reached where the demand for bonds just equals their supply. Therefore, the demand for speculative money balances – by bears – increases as the prices of bonds rise, or conversely, as the interest rate falls, so that the aggregate speculative demand for money is inversely related to the rate of interest.

Modern monetary and macroeconomic theory has abandoned this line of reasoning and has instead opted for an analysis based on portfolio selection, so that a better name for the money demand derived from portfolio selection analysis would be the "*portfolio demand for money*." This approach is presented in Chapter 5.

Tobin's formalization of Keynes's speculative money demand for an individual

Tobin's (1958)[21] formalization of Keynes's speculative demand analysis has become a classic and is presented in the following.

As with Keynes's analysis, Tobin assumes that there are only two assets, money and bonds, in which the individual can invest the amount of funds in his portfolio. Money is assumed to have a known yield of zero and is therefore riskless in the sense of possessing a zero standard deviation of yield. The bond is a consol, also known as a "perpetuity" in the United States, and has the characteristic that it does not have a redemption date, so that the issuer need never redeem it but may continue to make the coupon payment on it indefinitely.

In perfect capital markets, the market price of a consol will equal its present discounted value. Therefore, the price p_b of a consol which has a nominal coupon payment c per period, and is discounted at a market rate of interest x on loans, is given by:[22]

$$p_b = \frac{c}{1+x} + \frac{c}{(1+x)^2} + \cdots$$

$$= c\left(\sum_{t=1}^{\infty} \frac{1}{(1+x)^t}\right)$$

$$= c\left(\frac{1}{x}\right) = \frac{c}{x}$$

Therefore, the consol's value will equal its coupon rate (in perpetuity) divided by the market rate. For a given coupon value, an increase in the market interest rate will reduce the consol's price and imply a capital loss. In the special case of a bond that has the same coupon rate as the market discount rate, $c = x$, so that its market value will equal unity, i.e. $p_b = 1$.

[20] Parts of the analysis of this article are presented later in Chapter 5 on the speculative demand for money.

[21] The proof uses the mathematical formula that, for $x > 0$,

$$\sum_{t=1}^{\infty} \frac{1}{(1+x)^t} = \frac{1}{x}$$

Many bonds have a finite redemption date, say n. For this, the relevant formula is

$$\sum_{t=1}^{n} D^t = \sum_{t=0}^{n} D^t - 1 = \frac{1 - D^{n+1}}{1 - D} - 1$$

where $D = 1/(1 + x)$. Since $x > 0$, $D^{n+1} \to 0$ as $t \to \infty$.

Now assume that the market interest rate is \$$R$ per year and the consol is expected to pay a coupon \$$R$ per year in perpetuity. With a coupon of R in perpetuity, the above mathematical formula implies that the consol's present value at the market interest rate R would equal R/R and be 1.

Assume for the following analysis that the coupon payment on the consol is set at R and its current price is one dollar. Further, assume that the individual expects the market rate of return on consols to be R^e for the future, with this expectation held with a probability of one and independent of the current yield R. With R treated as the coupon payment and the rate of discount expected to be R^e in perpetuity, the expected value of the consol next year will be R/R^e. Therefore, the expected capital gain or loss G on the consol will be:

$$G = R/R^e - 1$$

The expected yield (R G) from holding a consol costing \$1 is the sum of its coupon R and its capital gain G. This sum is given by:

$$R + G = R + R/R^e - 1$$

If the yield (R G) were greater than zero, the rational individual would buy only consols, since they would then have a yield greater than money, which was assumed above to have a zero yield.[23] Conversely, if the yield on consols were negative, the individual would hold only money since money would be the asset with the higher yield.

The switch from holding bonds to money occurs at $R + G = 0$. This condition can be used to derive the *critical level* R^c of the current return R such that:

$$R^c = R^c/R^e - 1 = 0$$

which implies that:

$$R^c - R^e/\overset{\Sigma}{1} + R^e \quad \Sigma$$

For a given R^e, if the current interest rate R is above R^c, (R G) > 0 and only consols will be bought; if it is below R^c, (R G) < 0 and only money will be held. Therefore, in Figure 2.1, the individual's demand for money is the discontinuous step function (AB, CW); above R^c, the rational *individual's* whole portfolio W is held in consols and the demand for money along AB is zero; below R^c, all of W is held in money balances and the demand function is CW.

Keynes's overall speculative demand function

Keynes had argued that the bond market has a large number of investors who differ in their expectations such that the lower the rate of interest, the greater will be the number of investors who expect it to rise, and vice versa. Therefore, at high rates of interest, more investors will

R

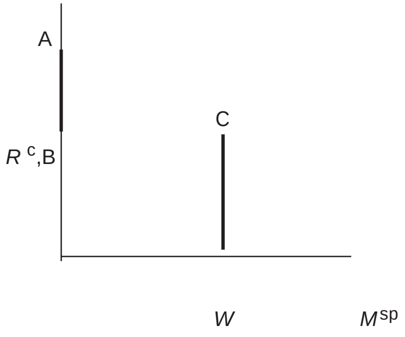


Figure 2.1

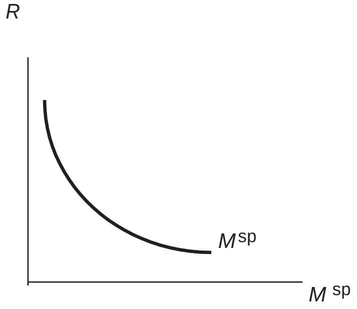


Figure 2.2

expect the rate to fall and few will hold money. At a somewhat lower rate of interest, a smaller number of the investors will expect the interest rate to fall and more of them will hold money. Hence, the aggregate demand for money will rise as the interest rate falls, and is shown as the continuous downward sloping curve M^{sp} in Figure 2.2. Therefore, Keynes's analysis implies that the speculative demand for money depends inversely upon the rate of interest, so that the speculative demand function for money can be written as:

$$M^{sp} = L(R) \tag{19}$$

where:

M^{sp} = speculative demand for money

R = market/nominal rate of interest.

Keynes called the function $L(R)$ the degree of *liquidity preference,* with L standing for liquidity.

Note that in Keynes's analysis, the individual allocates his financial wealth FW between money and bonds. Hence, in addition to the interest rate, financial wealth FW must be one of the determinants of their demand. Therefore, (19) needs to be modified to:

$$M^{sp} = L\,(R,\ FW\,)$$

There also exists the possibility that the economy could substitute among money, bonds and commodities as stores of value.[24] The analysis allowing this possibility would need to broaden the relevant wealth variable to total wealth and also make the speculative money demand function a function of both the return on bonds and that on commodities.[25] This extension of Keynes's analysis leads to Friedman's money demand function in the next section. For the time being, we continue with (19) for Keynes's specification of the speculative demand for money, thereby simplifying it by ignoring wealth as a determinant of the speculative demand for bonds.

Keynes's overall demand for money

Keynes argued that:

> Money held for each of the three purposes forms … a single pool, which the holder is under no necessity to segregate into three watertight compartments; for they need not be sharply divided even in his own mind, and the same sum can be held primarily for one purpose and secondarily for another. Thus we can – equally well, and, perhaps, better – consider the individual's aggregate demand for money in given circumstances *as a single decision, though the composite result of a number of different motives*.
>
> (Keynes, 1936, p. 195; italics added).

Hence, the aggregate demand for money, M, depends positively upon the level of income Y due to the transactions and precautionary motives and negatively upon the rate of interest R due to the speculative motive. In symbols,

$$M^{\mathrm{d}} = M^{\mathrm{tr}} + M^{\mathrm{sp}} = M(Y, R)$$

However,

> whilst the amount of cash which an individual decides to hold to satisfy the transactions-motive and the precautionary-motive is not entirely independent of what he is holding to satisfy the speculative motive, it is a safe first approximation to regard the amounts of these two sets of cash-holdings as being largely independent of one another.
>
> (Keynes, 1936, p. 199).

Hence, *as an approximation*, the demand function for money balances M^{d} is given by:

$$\begin{aligned} M^{\mathrm{d}} &= M^{\mathrm{tr}} + M^{\mathrm{sp}} \\ &= kY + L(R) = kPy + L(R) \end{aligned} \tag{20}$$

where $k > 0$ and $L(R) < 0$.

Liquidity trap

Keynes argued that the speculative demand for money would become "absolute" (infinitely elastic) at that rate of interest at which the bond market participants would prefer holding money to bonds, so that they would be willing to sell rather than buy bonds at the existing bond prices. Following Keynes's reasoning, the liquidity trap occurs at the rate of interest at which a generally unanimous opinion comes into being that the rate of interest will not fall further but may rise. At this rate, there would be a general opinion that bond prices will not rise but could fall, thereby causing capital losses to bondholders, with the existing rate of interest merely compensating for the risk of such a capital loss. In such circumstances, the public would be willing to sell all its bond holdings for money balances at their existing prices, so that the monetary authorities could buy any amount of the bonds from the public and, conversely, increase the money holdings of the public by any amount, at the existing bond prices and rate of interest. Therefore, once the economy is in the liquidity trap, the monetary authorities cannot use increases in the money supply to lower the interest rate.

As against this analytical presentation of the liquidity trap, Keynes asserted that "whilst this limiting case might become practically important in future, I know of no example of it hitherto" (Keynes, 1936, p. 207). Note that this assertion was made in the midst of the most severe depression in Western history; if the liquidity trap did not exist then, it can hardly have existed in more normal periods of economic activity. Hence, in Keynes's view, while the liquidity trap is an intellectual curiosity for monetary economics, it is not of practical relevance. Consequently, contrary to some expositions or critiques of Keynesian economics in earlier decades, Keynes did not build his macroeconomic model on the assumption of the liquidity trap.

Keynes's statement on the empirical non-existence of the liquidity trap is strictly incorrect under his own analysis of the speculative demand for money. In this analysis, the liquidity trap will come into existence whenever the dominant opinion in the bond market is that the market interest rates are going to rise, not decline. Such an opinion does quite frequently come into existence in the bond markets, so that the liquidity trap is not unknown in them. Further, such an opinion can exist at any level of the rate of interest and not merely at low or even single-digit rates. Furthermore, the liquidity trap will continue to exist until the dominant market opinion changes to envision possible decreases in the rate of interest.[26] This would happen once the interest rates have adjusted to the market opinion, so that the liquidity trap would usually exist for short periods, which may not be long enough to affect investment and the macroeconomy. Therefore, while liquidity traps may often come into existence in the normal day-to-day functioning of bond markets, their existence for macroeconomics could be quite insignificant.

In contrast to Keynes's reasoning, which emphasized the possibility of a capital gain or loss on holding bonds, for the existence of a liquidity trap, is the argument that nominal interest rates close to zero do not compensate individuals for the hassle and inconvenience of holding bonds when they could forgo these by holding money balances. This argument is supported by the analysis of the transactions demand for money presented in Chapter 4, where it is shown that, at low enough interest rates relative to the brokerage costs of conversion between bonds and money, it is not profitable to hold bonds, so that only money will be

[23] This, of course, is difficult to envision if the interest rate is already zero.

held; thus, in this low enough range of interest rates, the interest elasticity of money demand will be zero. In recent years, Japan is among the very few countries that have had short-term interest rates close to zero. Some empirical studies do report that the interest elasticity of money demand is much higher during Japan's low interest rate period than in other periods (see, for example, Bae *et al.*, 2006).

Keynes's and the early Keynesians' preference for fiscal versus monetary policy

Volatility of money demand

Keynes's analysis of the speculative demand for money made it a function of the subjective expectations of the bulls and bears in the bond and stock markets. Such expectations were quite volatile in the 1930s and can be quite volatile even nowadays, as one can observe in the day-to-day volatility of the stock markets and the periodic "collapse" or sharp run-ups of prices in them. Given this volatility, Keynes asserted that the speculative demand function for money was very volatile – that is, this function shifts often. Since Keynes believed that the speculative demand for money was a significant part of the overall demand for money, the latter would also be quite volatile. This would introduce a considerable degree of instability into the aggregate demand, prices and output in the economy, and also make the pursuit of monetary policy, which could trigger changes in the investors' expectations, very risky. Keynes, therefore, was more supportive of fiscal policy than of monetary policy as the major stabilization policy in the economy. It was also the general attitude of the Keynesians until the late 1950s.

Radcliffe report: money as one liquid asset among many

The early (1940s and 1950s) Keynesians' preference for fiscal policy as against monetary policy was reinforced by the Radcliffe report[27] in Britain in 1958, which argued that money was one liquid asset among many, of which trade credit was a major part, and that the economy was "awash in liquidity," so that changes in the money supply could not be used as an effective policy tool for changing aggregate demand in the economy. Therefore, Keynes's belief in the unreliability of the effects of monetary policy (because of the instability of the money demand function) was buttressed for the 1950s Keynesians by the Radcliffe report that the money supply was only a small part of the total supply of liquidity, which was the proper determinant of aggregate demand but could not be significantly changed by monetary policy.[28]

Given the above views on the impotence of monetary policy or the unreliability of its impact, Keynesians from the 1940s to the 1960s placed their emphasis for the management of aggregate demand on fiscal policy. They advocated the active use of fiscal deficits and surpluses for ensuring the aggregate demand needed to achieve a high level of output and employment.

Both of the above arguments against the use of monetary policy were discarded during the 1960s when, prodded by Friedman's views and empirical findings on the money demand function, the Keynesians and the neoclassicists – and later the 1970s monetarists – came to the conclusion that monetary policy, at that time interpreted as changes in the money supply, had a strong impact on the economy. Part of this achievement was due to the contributions of Milton Friedman.

Friedman's contributions

Friedman made profound contributions to monetary and macroeconomics, especially to the role of monetary policy in the economy. He believed that monetary policy had a strong impact on output and employment, but with a long and variable lag. Among his numerous contributions was his classic article on the "restatement" of the quantity theory.

Friedman's "restatement" of the quantity theory of money

Milton Friedman (1956), in his article "The quantity theory of money – a restatement," sought to shift the focus of the quantity theory and bring it into closer proximity with the developments in monetary theory up to the mid-1950s. Three strands of these developments are important to note. One development was that of Keynesian macroeconomics, which placed the determination of the price level in a broad-based macroeconomic model with product, money and labor markets, and restricted the analysis of the money market to the specification of demand, supply and equilibrium in the money market. This development had argued that the price level could be affected by shifts in the aggregate demand for commodities and that changes in the money supply could affect output, and not merely prices, in an economy operating at less than full employment. The second development was Keynes's emphasis on the speculative demand for money and therefore on the role of money as a temporary store of value for the individual's wealth. The third development was the integration of the theory of the demand for money into that of goods generally by treating money as a consumer good in the consumer's utility function and as an input in the firm's production function (Patinkin, 1965).

Friedman argued that the quantity theory was merely the proposition that *money matters*, not the more specific statement that changes in it will cause proportional changes in the price level. By "money matters," Friedman meant that changes in the money supply could cause changes in nominal variables and sometimes even in real ones, such as output and employment.

Friedman restated the quantity theory to limit its main role to that of a theory of the demand for money. For consumers, the demand for real money balances was made identical to that of other consumer goods, with real balances being one of the goods in the consumer's utility function. In this role, Friedman viewed real balances as an asset, with the real values of money, stocks, bonds and physical assets being alternative forms of holding wealth and incorporated into the individual's utility function. For firms, real balances were a durable good, similar to physical capital, with both appearing as inputs in the production function. Friedman, therefore, concluded that the analysis of the demand for money was a special topic in the theory of the demand for consumer and capital goods.

Further, Friedman argued that a unit of money is not desired for its own sake but for its purchasing power over goods, so that it is a good in terms of its real and not its nominal value. This real purchasing power of money over commodities is reduced by inflation, so that the rate of inflation is the opportunity cost of holding real balances as against holding commodities. Hence, money demand depends on the (expected) inflation rate.

Since money acts as a store of value, it is like other assets and its demand must also depend on the yield on other assets. These yields, to reflect the concern of the individual with his purchasing power, must be taken to be in their real and not their nominal value. Thus, in periods of inflation, the individual would discount the nominal yields on assets by the rate of inflation.

Friedman further argued, as in his consumption theory (the permanent income hypothesis of consumption), that the individual will allocate his lifetime wealth over commodities and over the liquidity services of real balances. This lifetime wealth (w) is the sum of the individual's human and non-human wealth, where human wealth (HW) is defined as the present discounted value of labor income while non-human wealth (NHW) consists of the individual's financial and physical assets. Since the value of these assets is known in the present, while future labor income is uncertain, the degrees of uncertainty affecting human and non-human wealth are quite different, so that their effects on the demands for commodities and money would also be different. Friedman proxied the individual's degree of uncertainty of wealth by the ratio of his human to non-human wealth.

Therefore, according to Friedman, the main determinants of the individual's demand for real balances were the *real* yields on other assets (bonds, equities and physical assets), the rate of inflation, real wealth and the ratio of human to non-human wealth. Writing this demand function in symbols,

$$m^{\mathrm{d}} = M^{\mathrm{d}}/P = m^{\mathrm{d}}(r_1,\dots, r_n, \pi, w, HW/NHW) \tag{21}$$

where:

m^{d} = demand for money balances in real terms
M^{d} = demand for money balances in nominal terms
P = price level
r_i = yield in real terms on the ith asset
π = rate of inflation
w = wealth in real terms
HW/NHW ratio of human to non-human wealth.

Permanent income as the scale determinant of money demand

Since data on human and total wealth was not available, Friedman proxied total wealth by permanent income y^{p}. At the theoretical level, the relationship between these variables is specified by:

$$y^{\mathrm{p}} = rw \tag{22}$$

where r is the expected average real interest rate over the future. Permanent income y^{p} can be interpreted as the average expected real income over the future. In line with Friedman's work on the consumption function, Friedman employed adaptive expectations – which use a geometric lag of past incomes – to estimate y^{p}, rather than rational expectations. These procedures will be covered in Chapter 8.

Since the demand function is derived from the consumer's utility function, which represents the individual's tastes, shifts in these tastes will shift the demand function. Friedman sought to take account of such shifts by incorporating a variable *u* for "tastes/preferences" in the demand function. Substituting y^p for *w*, taking *r* to be proxied by the various interest rates and adding the new variable *u* for tastes/preferences, in the manner of Friedman's article, the demand function for real balances becomes:

$$m^d = M^d/P = m^d(r_1, \ldots, r_n, \pi, y^p, HW/NHW, u) \quad (23)$$

Note that this demand for money is essentially derived from the notion of money as an asset – that is, a store of value – and that permanent income appears in it as a proxy for wealth.

Friedman on the velocity of money

Since the velocity of circulation *V* equals *Y* /*M* , and *M* in equilibrium equals M^d, we have:

$$V = \frac{y}{m^d(r_1, \ldots, r_n, \pi, y^p, HW/NHW, u)} \quad (24)$$

where both the numerator and the denominator on the right-hand side of the equation are real variables, so that their ratio is also a real variable. The preceding equation implies that, for Friedman, velocity was not a constant but a real variable, which depended upon the real yields on alternative assets and other variables. Except for the introduction of permanent income instead of current income as a determinant on the right side, (24) was consistent with the Keynesian tradition. The essential difference between Friedman and Keynes was on the stability of the velocity function: Friedman asserted that velocity was a function of a few variables and the velocity function was stable, whereas, for Keynes, the velocity function possessed, by virtue of the volatile nature of the subjective probabilities on bond returns, the potential for being unstable and its shifts unpredictable.

Friedman on the money supply

On the money supply, Friedman asserted that the supply function of money was independent of the money demand function. Further, some of the important determinants of the former, including political and psychological factors, were not in the latter. Hence, the money demand and supply functions were separate and could be identified in the data.

Friedman, like Keynes, assumed that the central bank determines the money supply, so that it could be treated as an exogenous variable for the macroeconomic analysis of the macroeconomy. This is, of course, a practical question. Its validity depends on central bank behavior. By the mid-1990s, many central banks were using the interest rate as their primary monetary policy instrument, while leaving the money supply to be determined endogenously by the economy at the set interest rate. While the exogeneity of the money supply was an unquestioned mainstay of short-run macroeconomic models and of the IS–LM analyses until the mid-1990s, the new Keynesian models which have emerged since the mid-1990s tend to assume that the central bank sets the interest rate, so that the money supply becomes endogenous in these models.

Friedman on inflation, neutrality of money and monetary policy

On the basis of his empirical studies, Friedman asserted that inflation is always and everywhere a monetary phenomenon. This assertion has become quite famous. While it does not accurately explain the determination of low inflation rates (i.e. in the low single digits), it does explain quite well persistently high inflation rates over long inflationary periods. The attribution of persistently high inflation rates to high money supply growth rates has already been explained in the earlier presentation of the quantity equation.

Friedman held that money was neutral in the long run. But, for the short term, he was strongly of the view that money was not neutral and, in fact, offered very significant and convincing economic evidence from the history of the United States that it was not so (Friedman and Schwartz, 1963, esp. pp. 407–19, 712–14, 739–40; Friedman, 1958). He also distinguished between anticipated and unanticipated changes in inflation rate and argued that the initial effects of a higher unanticipated inflation rate last for about two to five years, after which the initial effects start to reverse, so that the effects of unanticipated money supply and inflation increases on output, employment and real interest rates could last ten years (Friedman, 1968). To Friedman, changes in the money supply had a strong impact on output and unemployment,[29] and major depressions and recessions were often associated with severe monetary contractions.[30] Conversely, for the USA, major inflations were usually associated with wars,[31] during which the large fiscal deficits were financed by increases in the money supply.

However, Friedman maintained and showed that the timing of the impact of money supply changes on output was unpredictable and the lags involved were long and variable (Friedman, 1958). He concluded that major instability in the United States has been produced or, at the very least, greatly intensified by monetary instability. Consequently, he maintained that discretionary monetary policy could have unpredictable results and should not be followed. He claimed that:

> The first and most important lesson that history teaches us … is that monetary policy can prevent money from itself being a major source of economic fluctuations … [It] should avoid sharp swings in policy. In the past, monetary authorities have on occasion moved in the wrong direction. … More frequently, they have moved in the right direction, albeit often too late, but have erred by moving too far. … My own prescription … is that [it adopt] a steady rate of growth in a specified monetary total. … The precise rate of growth, like the precise monetary total, is less important than the adoption of some stated and known rate.
>
> (Friedman, 1968, pp. 12–16).

Friedman versus Keynes on money demand

Friedman's main concern in deriving his demand function was with money as a real asset held as an alternative to other forms of holding wealth, whereas Keynes's analysis was for the demand for nominal money balances. Friedman's analysis also implied that money demand depends on wealth or permanent income, rather than on current income as in Keynes's analysis. However, Friedman believed that the demand for money does not in practice become

[24] Friedman shared these views with most pre-Keynesian (traditional classical) economists.

[25] In the USA, during the Great Depression years 1929 to 1933, the money supply decreased by more than one- fourth, due to increases in the ratios of the public's currency holdings and of bank reserves to money supply, as well as bank failures.

[26] For the USA, this was so for the Civil War, World Wars I and II, Korean War and Vietnam War.

infinitely elastic, thereby agreeing with Keynes on the absence of the liquidity trap in practice. Further, Friedman believed that the money demand function was stable, whereas Keynes had adduced the subjective nature of probabilities in the absence of complete information on the future returns on bonds to derive the volatility of the speculative and overall money demand. On this point, Friedman's own and others' empirical findings for the 1950s and 1960s data supported Friedman over Keynes on the stability of the money-demand function (see Chapter 9).

Friedman further asserted that the money-demand and velocity functions were even more stable than the consumption function.[32] Till the late 1960s, the stability of the latter was the linchpin of the Keynesian analysis in its enthusiastic support for fiscal policy over monetary policy. Friedman's assertion meant that monetary policy would, at least, also have a strong impact on the economy. The success of Friedman's agenda was such that by the early 1960s the Keynesians had accepted monetary policy as having a strong and fairly reliable impact on aggregate demand, so that a synthesis – known as the neoclassical-Keynesian synthesis – emerged in the 1960s. This synthesis was reflected in the common usage of the IS–LM model for the macroeconomic analysis of the impact of monetary policy on aggregate demand. The divisions among these schools were henceforth confined to questions of the further impact of aggregate demand changes on output and unemployment.

At a general level, Friedman's money-demand analysis was not an elaboration or restatement of the quantity theory, despite Friedman's claim for it, and could more appropriately, as Patinkin (1969) pointed out, have been labeled a statement of the Keynesian money demand function or of the portfolio approach – as in Tobin (1958) – to money demand topical in the 1950s.[33]

Friedman was essentially a Keynesian in his macroeconomic theory and on his theory of money demand, but he was a conservative on the pursuit of monetary policy (Patinkin, 1981). On macroeconomics, his theoretical and empirical contributions showed that changes in the money supply could have strong effects on both nominal and real output. On monetary policy, Friedman advocated that an active monetary policy should not be pursued. Part of this advocacy was based on his roots in political conservatism and partly on his empirical finding that money supply changes impact on the economy with a long and variable lag

UNIT –II

Money Supply and Banking Institutions

Keynes had designated the transactions demand for money as due to the transactions motive but had not provided a theory for its determination. In particular, he had assumed that this demand depended linearly on current income but did not depend on interest rates.

Subsequent contributions by Baumol and Tobin in the 1950s established the theory of the transactions demand for money. These contributions showed that this demand depends not only on income but also on the interest rate on bonds. Further, there are economies of scale in money holdings.

The transactions demand for money is derived under the assumption of certainty of the yields on bonds, as well as of the amounts and time patterns of income and expenditures.

Key concepts introduced in this chapter

- ♦ Transactions demand for money
- ♦ Economies of scale in money demand
- ♦ Elasticity of the demand for real balances with respect to the price level
- ♦ Elasticity of the demand for real balances with respect to income
- ♦ Elasticity of the demand for real balances with respect to the interest rate
- ♦ Elasticity of the demand for real balances with respect to their user cost
- ♦ Efficient funds management

This chapter presents the main elements of the theory of the demand for transactions balances. In doing so, it follows Keynes in assuming that an individual's money holdings can be validly subdivided into several components, one of which is purely for meeting transactions.

Chapter 2 has pointed out that many of the classical economists and Keynes had made the simple assumption of the unit elasticity of demand for transactions balances with respect to nominal income. In particular, the demand for transactions balances was taken to double if either the price level or real income/expenditures – but not both – doubled. Hardly any analysis was presented in support of such a statement and it remained very much in the realm of an assumption.

Developments during the 1950s analyzed the demand for transactions balances rigorously from the standpoint of an individual who minimizes the costs of financing transactions by holding money balances and bonds, defined as interest-paying non-monetary financial assets. This analysis showed that the transactions demand for money depends negatively upon the bond rate of interest and that its elasticity with respect to the real level of expenditures is less than unity. The original analyses along these lines were presented by Baumol (1952) and Tobin (1956). The following presentation draws heavily upon Baumol's treatment of the subject.

Developments since the 1950s have extended and broadened the Baumol–Tobin transactions demand analysis, without rejecting it. The most significant extension of this analysis has been to the case where there is uncertainty in the timings of the receipts and payments. The demand for money under this type of uncertainty is usually labeled as the precautionary demand for money and is the subject of Chapter 6.

The basic inventory analysis of the transactions demand for money

This section presents Baumol's (1952) version of the inventory analysis of the transactions demand for money. This analysis considers the choice between two assets, "money" and "bonds," whose discriminating characteristic is that money serves as the medium for payments in the purchase of commodities whereas bonds do not; hence, commodities trade against money, not against bonds. There is no uncertainty in the model, so the yield on bonds is known with certainty. The real-world counterpart of such bonds is interest-paying savings deposits or such riskless short-term financial assets as Treasury bills. Longer-term bonds whose yield is uncertain are not really considered in Baumol's analysis. Baumol's other assumptions are:

1 Money holdings do not pay interest. Bond holdings do so at the nominal rate R. There are no own-service costs of holding money or bonds, but there are transfer costs from one to the other, as outlined later. Bonds can be savings deposits or other financial assets.

2 There is no uncertainty even in the timing or amount of the individual's receipts and expenditures.

3 The individual intends to finance an amount $\$Y$ of expenditures, which occur in a steady stream through the given period, and already possesses the funds to meet these expenditures. Since money is the medium of payments in the model, all payments are made in money.

4 The individual intends to cash bonds in lots of $\$W$ spaced evenly through the period. For every withdrawal, he incurs a "brokerage (bonds–money transfer) cost" that has two components: a fixed cost of $\$B_0$ and a variable cost of B_1 per dollar withdrawn. Examples of such brokerage costs are broker's commission, banking charges and own (or personal) costs in terms of time and convenience for withdrawals from bonds. The

overall cost per withdrawal of $\$W$ is $\$(B_0 + B_1W)$.

Since the individual starts with $\$Y$ and spends it in a continuous even stream over the period, his average holdings, over the period, of the funds held in bonds B and money M are only $Y/2$. Hence, $M+B = ½Y$.[1] Further, since the individual withdraws W each time and spends it in a continuous steady stream, and draws out a similar amount the moment it is spent, his average transactions balances M are $½W$. These propositions are shown in Figures

and 4.2. In Figure 4.1, for expenditures over one period, the triangle 0Y1 represents the amount of income that has not been spent at the various points of time within the period and 1YA is the amount that has been spent. 0Y1 equals $½Y$ over the period and would be held

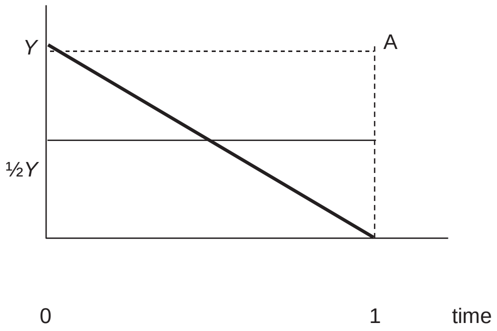


Figure 4.1

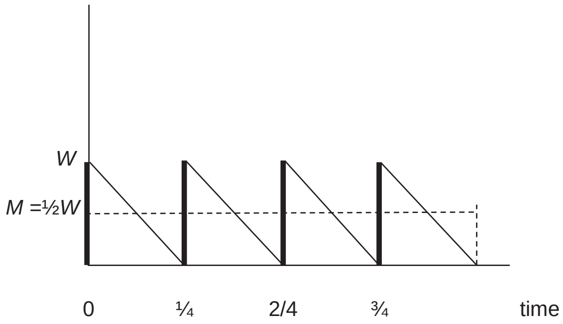


Figure 4.2

in either money or bonds. Figure 4.2 focuses on money holdings. To illustrate, assuming that the period is divided into 4 weeks, the amount $\$W$ is withdrawn at the beginning of each week and spent evenly through the week. The average money balances over the period are only $\tfrac{1}{2}W$, and, from Figures 4.1 and 4.2, the average bond holdings over the period are $(\tfrac{1}{2}Y - \tfrac{1}{2}W)$.[2]

Since the total expenditures of Y are withdrawn from bonds in lots of W, the number n of withdrawals is (Y/W). The cost of withdrawing Y from bonds is the cost per withdrawal times the number of withdrawals and is given by $[(B_0 + B_1 W)n]$. In addition, the interest foregone by holding money rather than bonds is RM. Since $M = \tfrac{1}{2}W$, this interest cost equals $RW/2$. The total opportunity cost C of financing Y of expenditures in this manner is the sum of the cost of withdrawing Y from investments and the interest foregone in holding average money balances of $(W/2)$. Hence,

$$\begin{aligned} C &= RM + (B_0 + B_1)n \\ &= RW/2 + B_0 \cdot Y/W + B_1 Y \end{aligned} \tag{1}$$

If the individual acts rationally in trying to meet his payments Y at minimum cost, he will minimize the cost C of holding transactions balances. To do so, set the derivative of (1) with respect to *W* equal to zero. This yields:

$$\partial C/\partial W = R/2 - B_0 \cdot Y/W^2 = 0 \qquad (2)$$

so that:

$$W = \Sigma 2B_0 \cdot \ Y/R\ \Sigma^{\frac{1}{2}} \qquad (3)$$

and

$$M^{\mathrm{tr}} = \tfrac{1}{2}W = (\tfrac{1}{2}B_0)^{\frac{1}{2}}Y^{-\frac{1}{2}}R^{-\frac{1}{2}} \qquad (4)$$

where we have inserted the superscript tr to emphasize that (4) specifies only the transactions demand for money and does not include the money demand that would arise for speculative and other motives. (4) is called the *square root formula* in inventory analysis and has the easily identifiable form of a Cobb–Douglas function. In the present analysis, it specifies the demand for transactions balances for a cost-minimizing individual. The preceding demand function is clearly different from Keynes's demand function for transactions balances and, among other things, indicates that *the demand for transactions balances depends upon the nominal rate of interest*. The properties of this demand function, showing its response to changes in the real levels of expenditures, interest rates and prices, are discussed below.

Brokerage costs are the prices charged for brokerage services, which are commodities (i.e. "goods and services"), so that: let $B_0 = P.b_0$ and $B_1 = P.b_1$, where b_0 and b_1 are the elements of the brokerage charge *in real terms*, whereas B_0 and B_1 were nominal brokerage charges, and *P* is the price level. The reason for expressing brokerage costs in this way is that the brokerage services related to money withdrawals from earning assets are themselves commodities and, from a rigorous viewpoint, if the prices of *all* commodities double, the brokerage cost must also double. Hence, both B_0 and B_1 must be taken to increase in the same proportion as the commodity price level *P*.

Therefore, equation (4) can be rewritten as:

$$M^{\mathrm{tr,d}} = (\tfrac{1}{2}b_0)^{\frac{1}{2}}Py^{\frac{1}{2}}R^{-\frac{1}{2}} \qquad (5)$$

and

$$M^{\ \mathrm{tr,d}}/P = m^{\mathrm{tr}} = (\tfrac{1}{2}b_0)^{\frac{1}{2}}y^{\frac{1}{2}}R^{-\frac{1}{2}}$$

Therefore, the elasticities of the transactions demand for money with respect to *y*, *R* and *P* are:[3]

$E_{m.y} = \tfrac{1}{2}\ E_{m.R}$

$= -\tfrac{1}{2}\ E_{M.P} =$

1

$E_{m.P} = 0$

In (5), since the elasticity of demand for real transactions balances with respect to *real* income is only ½, the demand for real transactions balances increases less than proportionately with the individual's real income due to economies of scale in the cost of money withdrawals from bonds. The elasticity of the transactions demand for money with respect to the nominal interest rate is ½: the higher the interest rate, the higher is the cost of holding funds in transactions balances and the lower is the demand for such balances. The elasticity ($E_{m.P}$) of the transactions demand for *real* balances with respect to an increase in all prices is zero, consistent with that derived for the general demand for money in Chapter 3. By implication, from (5), the elasticity $E_{M.P}$ of the transactions demand for nominal balances is 1.

Elasticity of the demand for nominal balances with respect to nominal income

We can now refine the implications of this analysis for the elasticity ($E_{M.Y}$) of the demand for *nominal* balances with respect to *nominal* income Py. Intuitively, since $Y = Py$, nominal income changes if either real income y or prices P change. Consequently, at rates of inflation close to zero, $E_{M.Y}$ will be approximated by $E_{m.y}$, which is ½ in the above analysis. The higher the inflation rate, the more significant will be the influence of $E_{M.P}$, which is unity, so that in hyperinflation, $E_{M.Y}$ will approximate unity. Therefore, $E_{M.Y}$ will not be a constant over time but will vary between one-half and one, depending on real income growth relative to the inflation rate during the period under study. Both output and the price level change each period, so that, if their rates of change were roughly equal, Baumol's model implies that the estimated value of $E_{M.Y}$ should be near the mid-point of the range between 0.5 and 1.0. In fact, for developed economies with low rates of inflation, it is not unusual to find estimates of this elasticity somewhere near the middle (0.75) of the potential range. However, we should expect that economies with high (double-digit or higher) inflation rates would have higher estimated values of $E_{M.Y}$. In the limiting case of hyperinflation, the value of $E_{M.Y}$ should approach unity. These considerations imply that the estimated elasticity of demand for nominal transactions balances with respect to nominal income is likely to differ among sample periods if they have different growth rates of real output and prices.

Some special cases: the profitability of holding money and bonds for transactions

The above analysis incorporates the choice between holding money and the income-earning asset – "bonds" – to finance transactions. In exercising this choice, the individual will buy bonds only if he can make a profit from holding them; if he cannot, he will only hold money and equation (5) for the demand for real balances will not apply to him. For an analysis of this possibility, we need to derive the profit function from holding money and/or bonds.

As we have shown earlier through Figure 4.1, under Baumol's assumptions the individual spends his income Y in an even stream over the period and therefore holds ½Y on average in either money or bonds.[4] His average nominal holdings B of bonds are, therefore, equal to (½ $Y - M$), where, as before, M equals ½ W. The individual earns interest at the rate R on these bond holdings. The profit from holding either money or bonds equals this interest income from holding bonds less the brokerage cost of money withdrawals from bonds.[5] That is, the profit π from using the combinations of money and bonds is given by:

$$\pi = \text{interest income from bonds} - \text{brokerage expenses}$$

$$= R \cdot B - (B_0 + B_1 W)n$$

$$\cdot \qquad \Sigma \quad \cdot \qquad\qquad\qquad \Sigma$$

$$= R\,½Y - M \;-\; ½B_0 Y/M + B_1 Y \tag{6}$$

Maximizing (6) with respect to M yields the first-order maximizing condition as:

$$\partial\pi/\partial M = -R + ½B_0 Y/M^2 = 0 \tag{7}$$

Hence, as in (4),

$$M^{\text{tr}} = (½B_0)^{½} Y^{½} R^{-½} \tag{4}$$

Further,

$$B^{\text{tr}} = ½Y - (½B_0)^{½} Y^{½} R^{-½} \tag{8}$$

where the superscript tr on B emphasizes that this demand for bonds is only for transactions purposes. Hence, from (6),

$$\cdot \qquad\qquad\qquad \Sigma \quad \cdot \qquad\qquad\qquad\qquad \Sigma$$

$$\pi = R\,½\,Y - (½\,B_0)^{½} Y^{½} R^{-½} - (½B_0)Y/[(½\,B_0)^{½} Y^{½} R^{-½}] + B_1 Y$$

$$\cdot \qquad\qquad \Sigma$$

$$= ½\,RY - (½\,B_0)^{½} Y^{½} R^{½} - (½\,B_0)^{½} Y^{½} R^{½} - B_1 Y$$

$$= ½RY - 2(½B_0)^{½} Y^{½} R^{½} - B_1 Y \tag{9}$$

Simplifying, we get:

$$\pi = ½\,RY - 2RM - B_1 Y$$

$$= (½R - B_1)Y \;- 2RM \tag{10}$$

The last equation has an easy intuitive explanation: total interest income from holding money and bonds is reduced by the interest cost of holding money and the variable cost of withdrawing Y from bonds. Further, since the second term on the right-hand side is non-positive, the first term implies that, no matter what the level of income, it would not be profitable to hold bonds unless $R > 2B_1$.

In equation (10), π is non-positive if $R=0$ or if the total brokerage charges exceed the income from holding bonds. The latter would occur if the brokerage costs are relatively high. Note in this regard that the brokerage costs include both the charges explicitly levied by financial institutions and any other costs of conversion from bonds to money. The latter include the time and inconvenience, etc. – sometimes referred to as the "*shoe-leather costs*"– of trips to the banks and other relevant financial institutions. These costs can be quite high in

areas poorly served by financial institutions, as is common in the rural areas of developing economies and even sometimes of developed ones. They are dominant for individuals for whom the banks refuse to open accounts or for those who cannot meet the conditions – for example, acceptable references or minimum deposit balances – set by banks for opening or holding such accounts. In these cases, the individual will not find it profitable to hold bonds and will only hold money.

The profit from holding bonds in the transactions process is also non-positive if either income or/and interest rates are sufficiently low. Such considerations are relevant to relatively poor individuals or where the financial system and its regulation limits the interest rates that can be paid on bonds. In these cases, the individual's demand for bonds would again be zero.[6]

In cases of non-positive profits from holding bonds, i.e. $\pi \leq 0$, the demand for bonds would be zero and the optimal transactions demand for *nominal* balances would be:

$$M^{\text{tr}} = ½Y \qquad (11)$$

which has a unit income elasticity and a zero interest elasticity.

From (11), the transactions demand function for real balances m^{tr} is:

$$m^{\text{tr}} = ½y \qquad (12)$$

so that $E_{m.y}$ 1 and $E_{m.r}$ 0. In the nineteenth and early twentieth centuries the brokerage costs even for savings deposits were high,[7] while the incomes of most individuals were quite low, so that the money demand function was likely to be closer to (11) than to that implied by the inventory model. That is, the income elasticity of money demand was closer to unity, rather than to 0.5, and the interest elasticity was closer to zero, rather than to 0.5.

Even in the modern period, almost all economies have some individuals – usually those with lower incomes – with such a demand function. The more under-developed the financial system of the country or the local area, and the lower the incomes of the people, the more significant would be this factor. The inventory demand formula (5) thus tends to have limited validity for many less-developed economies and rural areas, and even for some segments of the population in the developed economies.

Demand for currency versus demand deposits

The above analysis does not really address the interesting question of the relative demands for currency, which is notes and coins, as against that for demand deposits. For this, we need to consider the cost, convenience and safety of holding and using currency as opposed to demand deposits in making payments, rather than the costs of conversion from "bonds" into these two forms of money. In the choice between using currency or demand deposits, demand deposits do have positive own costs of usage since they require some trips to the bank for making deposits and the banks often levy deposit and withdrawal charges on checks, whereas currency holdings do not involve any such charges for making payments from them.

Further, the most common types of demand deposits do not pay interest. Consequently, currency involves lower own costs of usage, so that the optimizing individual will hold currency only and not demand deposits. This seems to be the case in many less-developed economies and especially in those rural areas poorly served by banks.

However, it is patently not the case in most developed economies or the urban sectors of developing economies that most individuals do not hold demand deposits, so there must be other considerations which are relevant to the choice between currency and demand deposits. The major one here for most individuals seems to be the relative safety of holding demand deposits as against that of holding currency.[8] The concern with theft and robbery if large sums were kept or carried in currency was a major reason for the origin and spread of deposit banking in eighteenth- and nineteenth-century Europe and continues to be a major determinant of the relative demand for demand deposits versus currency. The greater the concern with the safety and convenience of currency holdings, the lower will be the relative demand for currency balances. To illustrate, Japan, with an extremely low theft and robbery rate, is an economy in which ordinary persons do not normally hold demand deposit accounts but pay for most transactions in money. Conversely, persons carrying large sums in currency in the United States would be very concerned about their personal safety and the safety of these sums, and tend to prefer to hold demand deposits for meeting most of their transactions needs.

Impact of economies of scale and income distribution

Distribution of income

Consider the following two cases:

(A) An economy with *n* individuals, but with one having the whole of the national income *Y* and the rest with zero income. With zero income, the latter do not find it profitable to hold bonds and also do not hold money.

(B) An economy with *n* identical individuals, each having an identical income *Y/n*.

From (5), the nominal demands for transactions balances are:
For (A):

$$M_A^{tr} = (\tfrac{1}{2}B_0)^{\frac{1}{2}}Y^{\frac{1}{2}}R^{-\frac{1}{2}} \qquad (13)$$

For (B):

$$M_B^{tr} = n^{-}\,\tfrac{1}{2}B_0)^{\frac{1}{2}}(Y/n)^{\frac{1}{2}}R^{-\frac{1}{2}}\Sigma$$

$$= n^{\frac{1}{2}}M_A^{tr} \qquad (14)$$

Since $n \geq 2$, $M_B^{tr} > M_A^{tr}$. Hence, the equal distribution of incomes leads to a higher demand for real balances.

A more realistic scenario than either (A) or (B) would be one where a certain number of poor individuals have positive incomes but do not find it profitable to hold bonds. Their income elasticity of money demand would be one. In this case, the economy would have two types of individuals, ones with the usual Baumol money demand function and the others with M ½ Y , so that the unequal distribution of incomes in this case produces greater money demand than under either (A) or (B). In the limiting case, imagine a scenario where incomes are equal but everyone is too poor to profitably hold bonds. This would imply the highest money demand, which would equal ½ Y.

Hence, provided all individuals have sufficient incomes to find it profitable to hold bonds, Baumol's model implies that the more unequal the distribution of incomes in the economy, the smaller will be the demand for real balances. However, this result may not hold if the unequal distribution of incomes leads to some individuals holding only money.

Economic development

The following analysis provides another example of the impact of economies of scale on the transactions holdings of money. Assume that, *ceteris paribus*, a fraction α of the population has enough income to hold transactions balances according to the inventory model, while the "poor" fraction (1 α) does not find it profitable to hold bonds for transactions purposes. The overall transactions demand for money per capita for the population is given by:

$$M^{\text{tr}} = \alpha(\tfrac{1}{2}B_0)^{\frac{1}{2}} Y_A^{\frac{1}{2}} R^{-\frac{1}{2}} + (1-\alpha)\cdot \tfrac{1}{2} Y_B \qquad (15)$$

where Y_A is the income of each better-off person and Y_B is the income of each poor one. These money holdings have an income elasticity between ½ and 1. As the brokerage cost B_0 declines due to financial development or/and as the income of each poor individual rises sufficiently, α rises. This would raise the interest elasticity of overall transactions holdings in the economy and reduce their income elasticity. Therefore, economic and financial development should lead to a decrease in income elasticity and an increase in interest elasticity.

Further, note our earlier result that for a given interest rate, if brokerage costs are high and incomes low, as they are in many developing countries and especially outside the big cities, it would be unprofitable to hold bonds, so that the income elasticity of money demand would be one and the interest elasticity would be zero. In such a context, the average elasticity of money demand would be closer to one than to ½, which is implied by Baumol's model. This result would be reinforced in a context of inflation, since the price elasticity of nominal money demand is one.

Efficient funds management by firms

The preceding analysis was couched in terms of an individual but it can also be applied to firms. In the case of a firm with many branches, is it optimal for the firm to have centralized or decentralized money management? Centralized money management is here taken to mean a single account held by the firm as a whole, with the central financial department treating all the branches as one unit for its decisions on the amounts to be withdrawn each time. The amount withdrawn is then allocated among the branches. Decentralized money management means separate accounts and separate decisions on the amounts to be withdrawn at any one time.

Consider the case of a firm with total income or receipts equal to $\$Y$ and having n identical branches, each with income/receipts equal to $\$Y/n$. If it has centralized funds management, with a single demand deposit account and investments from it into bonds, its cost-minimizing transactions balances would be as specified by (13). If it has decentralized funds management, with each branch holding its own demand deposit account and bonds, its transactions balances will be as specified by (14). The latter is larger the greater the number of branches.

Since centralized funds management implies lower transactions balances, it also implies higher profits. The efficient firm would, therefore, choose to centralize its fund management, all other things being the same. However, there are other factors that make at least partial decentralization of bank accounts desirable for firms. Among these are the convenience, bookkeeping and security aspects. Many firms consider these sufficiently significant to retain decentralized banking arrangements, with the balances being transmitted from the branches to a main account at periodical intervals or when they reach pre-specified levels. Hence, convenience and security reasons play important roles in the choice of the extent of centralization of deposits, as they do in the use of currency versus demand deposits.

In recent decades, the increasingly efficient electronic transfer and investment of funds have reduced brokerage costs and made it profitable for large firms to invest their surplus funds for periods as short as a day. Their desired end-of-the-day holdings of demand deposits may then be zero. Unpredictable withdrawals or deposits of funds can still occur, but these may be covered through overdraft facilities prearranged with the banks. In such a context, the actual holdings of demand deposits would be largely random. Such firms could still have positive currency demand but this would be largely in the nature of working or petty cash and depend upon considerations – for example, the unpredictable and uneven pattern of receipts and expenditures – other than those incorporated in the Baumol model.

The demand for money and the payment of interest on demand deposits

Many types of demand deposit accounts now pay interest. In order to properly consider these, assume that there are two assets, demand deposits and bonds, with each paying interest. Since currency does not pay interest, we exclude it from the definition of money in this section, so that money will equal demand deposits. The other assumptions of the model are as originally specified, including that the purchases of commodities can only be paid for by check drawn on a demand deposit account. As before, bonds are assumed to pay interest at the rate R, while demand deposits are now assumed to pay the rate R_D.

As in the preceding analysis, the average amount of demand deposits D is $W/2$ and that of bonds is $(½Y - D)$. The profit π from the use of money and bonds is:

$$\pi = R\left[½Y - D\right] + R_D D - \left[½B_0 Y/D + B_1 Y\right] \tag{16}$$

which yields the first-order maximizing condition as:

$$\partial\pi/\partial D = -R + R_D + ½B_0 Y/D^2 = 0 \tag{17}$$

Hence,

$$D^{\text{tr}} = (½B_0)^{½} Y^{½} (R - R_D)^{-½} \tag{18}$$

$$B = ½Y - (½B_0)^{½} Y^{½} (R - R_D)^{-½} \tag{19}$$

where:

$$E_{D.(R-R_D)} = -\tfrac{1}{2} \qquad \Sigma \tag{20}$$

$$E_{D.R} = -\tfrac{1}{2}\, R/(R - R_D) \tag{21}$$

The demand for transactions balances now depends upon the interest rate *differential* (R R_D), and the elasticity of the transactions demand for demand deposits with respect to the differential in the interest rates is ½. However, this elasticity with respect to the bond rate of interest alone – that is, if the bond rate rises but the interest rate on demand deposits stays unchanged – is now Σ ½ $R/(R$ $R_D)$: the higher the interest rate on demand deposits, the higher is the elasticity Σ of the demand for such balances. Since these elasticities are different, the impact on the demand for money of changes in bond yields will depend upon whether or not the interest rate on demand deposits is also changing.

Demand deposits versus savings deposits

As explained at the beginning of this chapter, non-checkable savings deposits can be viewed as a “bond” which pays interest but which cannot be directly used to make payments to others,[9] so that funds have to be transferred from savings accounts to checking accounts before a payment from them can be made by check. Prior to the advent of automatic banking machines and of telephonic and electronic transfers, a trip had to be made to a bank branch to transfer funds from savings accounts to a checking account or to obtain currency. Such a trip involved time and inconvenience, which are elements of the brokerage cost in the Baumol model. The proliferation of automatic banking machines and the general reduction in the banks’ conditions and charges for such transfers have reduced this element of brokerage cost very considerably. The electronic transfer of funds among accounts handled through one’s home computer has made this cost relatively insignificant.

Up to the 1960s, commercial banks also often imposed other costs, sometimes including a period of prior notice for withdrawal from savings accounts, for handling such transfers. The imposition of such notice has virtually disappeared. The result is that payments from savings deposits are now not very different in terms of costs and delays than from demand deposits.

For the following analysis, assume that savings deposits are the only kind of bond and the amount of savings deposits is designated as S. Since S replaces B in the analysis of section 4.2, the optimal ratio D/S in the context of that section is given by:

$$D/S = 1/\ \tfrac{1}{2}Y/D - 1 \quad \Sigma \tag{22}$$

$$= 1/\ \tfrac{1}{2}(\tfrac{1}{2}B_0)^{-\frac{1}{2}}Y^{\frac{1}{2}}R^{-\frac{1}{2}} - 1 \quad \Sigma \tag{23}$$

so that demand deposits fall with the decrease in brokerage costs. In the limit, $D/S \rightarrow 0$ as $B_0 \rightarrow 0$.

Historically, as the brokerage costs between demand deposits and savings accounts decreased, the proportion of balances held in demand deposits fell, so that this proportion is currently less than 10 percent of M2 in the United States and Canada. Increasing familiarity in the handling of transfers between bank accounts from telephones and home computers is likely to further reduce this proportion.

The proliferation of automatic banking machines has also reduced the brokerage costs of transfers between currency and demand deposits, and also between currency and savings accounts. Therefore, as implied by Baumol’s model, these banking facilities have allowed individuals to reduce their holdings of currency as against holding demand deposits and saving deposits. These banking facilities have therefore led to a decrease in both currency and demand deposits, so that the amounts held in M1 have fallen sharply.

Technical innovations and the demand for monetary assets

Recent decades have seen a considerable variety of innovations in the financial sector. The broad categories of these have been:

1 The creation of new types of financial assets and the increasing liquidity of some of the existing assets. These encompass institutional innovations such as interest-bearing demand deposits and checkable savings deposits, which did not become prevalent until the 1970s. They also include the issuance by banks of money market and other mutual funds, without a significant monetary brokerage charge for buying and selling such funds, and their divestiture into demand deposits at short notice. Such money market mutual funds, especially those sold by banks, became common only in the 1990s. Such innovations have shifted the transactions demand functions for currency, demand deposits and savings deposits.

2 Technical innovations in the deposit and withdrawal mechanisms and practices for various types of assets. These encompass the introduction of automatic teller machines (mainly in the 1980s) and telephonic and computer-based transfers of funds between accounts, beginning in the late 1990s but in common usage in this century. Debit cards are of this nature. They reduce the brokerage costs of using deposits, as against using checks, so that they reduce their transactions demand.

3 The development of “smart cards,” which store nominal amounts, just as a coin or banknote does, and which allow the transfer of all or part of this amount to others at the point of the transaction without involving a third party such as a bank or a credit card company. Examples of these are certain types of telephone cards. Leaving aside the differences in technology and focusing on the economic nature, such cards are similar to coins and notes, which also embody value and allow the transfer of the whole or part of this value by the bearer to another person, the transaction proceeding with anonymity with respect to other parties. A rather insignificant difference is that paying with a larger note than necessary involves a reverse payment of “change,” whereas the smart card allows transfer of the exact amount. The more important difference would be that a smart card with owner-authentication procedures built into it would prevent its theft to a much greater extent than is possible with currency, which can be used by the bearer without any authentication of proper ownership, so that the smart card would be more secure. This fea- ture should make smart cards more attractive, and their use could replace that of both cur- rency and checking accounts to a significant extent. In so far as both currency and smart cards constitute “value-carrying purses,” the former being a non-electronic one and the latter an electronic one, it would be appropriate to lump them together in the total

demand for “currency/purses” as against the demand for demand deposits, savings deposits, etc.

4 The development of digital cards, payments with which require the intervention of a third party such as a bank to verify, authorize and clear transactions over a network connection. These are more like checks or debit cards – whereas electronic purses are more like currency – and combine the advantages of checks with those of a credit or debit card. They leave a trail of transactions, which can be valuable for bookkeeping and security reasons.

5 The development of online payments, which allow payments to be made directly from a bank account to a payee. In the preceding analysis of the transactions demand for money, online payments reduce the monetary and non-monetary brokerage costs of using demand deposits and reduce their demand.

Hence, the very considerable – and continuing pace of – innovations in the financial industry in the past few decades have reduced the demand for currency, demand deposits and savings deposits, and have therefore shifted the demand functions for M1, M2 and the still wider definitions of money.

Estimating money demand

The inventory model of money demand implies that the alternative estimating log-linear forms of the transactions money demand equations are:

$$\ln M^{tr,d} = \beta_0 + \beta_y \ln y + \beta_R \ln R + \beta_P \ln P \tag{24}$$

$$\ln m^{tr,d} = \beta_0 + \beta_y \ln y + \beta_R \ln R \tag{25}$$

The model implies the elasticities $\beta_y = ½$, $\beta_R = -½$ and $\beta_P = 1$. But if the estimating equation had been formulated as:

$$\ln M^{tr,d} = \alpha_0 + \alpha_y \ln Y + \alpha_R \ln R \tag{26}$$

the estimate of α_Y should lie between ½ and 1, with this value being larger the greater is the inflation rate relative to the real output growth rate. Further, in economies in which income, interest rate and brokerage costs are such as to make it unprofitable for most of the public to hold bonds for transactions purposes, the real income and nominal income elasticities would be closer to unity.

A rise in the interest rate causes two effects. One, it induces individuals to trade more often between money and financial assets, as the above inventory demand model shows. Two, for some individuals, who did not formerly trade between money and financial assets because of the unprofitability of doing so, the rise in the interest rate makes it profitable to undertake such trades, thereby increasing the interest elasticity of transactions money demand for the population as a whole. Hence, this elasticity should be non-linear.[10] Similar considerations applied to increases in income from very low levels would also cause the income elasticity to be non-linear.

In applying the above inventory analysis to the data collected on money balances, note that while the theory specifies average money balances held, the data is often collected as end-of-day (or other period) data. Further, the financially developed economies, with electronic transfers of funds at virtually zero brokerage costs, usually have one-day and overnight loan markets, in which firms using efficient cash management procedures can invest their excess money balances at the end of the day and others short of funds can borrow them. For *sweep accounts*, the banks themselves monitor the state of their customers' accounts at the end of each day and *sweep* the accounts of excess balances, investing them in overnight money market funds. In such a case, the customers need to ensure that they keep only the minimum desired balances; any amounts above or below these are lent or borrowed in the overnight or day-to-day loan markets or through loan arrangements such as overdrafts with their own bank (Bar-Ilan, 1990). Therefore, for customers with large balances and low transactions costs, the desired minimum transactions balances at the end of the day would be zero under the simpler versions of the inventory analysis. In other models, the amounts held by firms would be random, with a zero mean.

In a more realistic context, the customer, often an individual rather than a firm, would hold positive balances but these would be determined by institutional arrangements such as the minimum compensating balances banks sometimes require their customers to maintain in lieu of banking charges. Such considerations and the inventory model are more applicable to households' rather than to large firms' transactions demand for money.

Note that the data on money balances does not differentiate between those held for transactions and those held for speculative or other purposes. Hence, the preceding transactions demand elasticities provide only rough guides to the overall money demand elasticities. Further, financial innovation in recent decades is likely to have shifted the money demand function, so that the estimated elasticities would differ among different sample periods. Numerous empirical studies (see Chapter 9) confirm this finding.

Empirical findings

At a general level, the Baumol/Tobin analysis of transaction demand implies that the interest elasticity of money demand in developed economies with developed financial sectors will be negative. This has now been confirmed beyond any doubt by empirical studies on the overall demand for money (see Chapter 9).

The preceding analyses of transaction demand imply that the income elasticity of real balances with respect to real income is ½, their interest elasticity is $-$½ and the price elasticity of nominal money balances is one. Further, if it is unprofitable for the individual to hold bonds, because incomes and interest rates are relatively low and brokerage costs relatively high, the income elasticity of real balances with respect to real income is one, their interest elasticity is zero and the price elasticity of nominal money balances is one. Therefore, the decision to hold transactions balances involves two decisions: (i) whether to hold non-monetary interest-bearing financial assets; and
(ii) how to allocate financial wealth between money and non-monetary interest-bearing financial assets. As income rises from low levels or brokerage costs fall with financial development, the average estimated income elasticity of real balances falls from a value close to one to a value close to ½. Empirical studies on the overall demand for money do usually estimate the income elasticity of money demand to be less than one (see Chapter 9).

For the usual income distributions with different incomes, the interest elasticity of transaction balances would be lower at low interest rates than at high interest rates, since more individuals in the population would find it profitable not to hold bonds, so that more of them would have zero interest elasticity of money demand. As interest rates rise, more and more individuals would find it profitable to hold some bonds and substitute between money and bonds, so that the interest elasticity would increase towards ½. Therefore, money demand will be non-linear with respect to the interest rate, as will be the interest elasticity of the transactions demand. Mulligan and Sala-i-Martin (2000), using a cross-section sample of countries, confirm such non-linearity for money demand as a whole, with low interest elasticity at low income levels. For developing economies, as incomes rise, more and more individuals will find it profitable to use banking services and switch between (non-interest yielding) money and interest-bearing assets, so that the interest elasticity of money demand should rise and the income elasticity of nominal money balances should fall.

Since the transactions demand for money is only a component of the overall money demand, which cannot be separated in empirical estimation into its components, the empirical findings on money demand are left for detailed consideration to Chapter 9.

Conclusions

The basic conclusion of the inventory analysis for transactions demand is that, assuming positive profits from holding some bonds (including savings deposits) as part of the transactions portfolio, households will have economies of scale in holding demand deposits, and a negative interest elasticity with respect to the interest rate. This elasticity will differ depending upon whether or not interest is paid on demand deposits and upon the interest rate differential.

Innovations in electronic transfers and centralized control between the head office and branches, and between firms' branches and banks, have reduced the inconvenience connected with centralization and have thus promoted greater centralization of money management. Further, they have reduced brokerage costs for firms. In the limiting case where the brokerage costs per transaction at the margin tend to zero, the demand for demand deposits would tend to zero. As a consequence, the transactions balances held by firms relative to their revenues have fallen. Variations in these balances may be largely dominated by random factors in the case of large firms with efficient funds management in well-developed financial markets.

A consideration that leads to positive demand deposits being held are minimum compensating balances, often in lieu of transactions fees, sometimes required by banks. Such banking practices, as well as the number and sizes of branches, would be among the major determining factors determining the minimum holdings of money balances by individuals and firms.

The above discussion implies that the aggregate demand for transactions balances in the economy has three components. These are:

1. The demand by households and firms who do not find fund management with some investment in interest-bearing non-monetary assets ("bonds") profitable and hold only money. Such a component will exist in virtually any economy but may only be a significant part of the whole in economies with undeveloped banking and other financial facilities or in developing countries with low average incomes.
2. The transactions balances of those households and firms that find such financial management profitable. For these, the Baumol model would be applicable.
3. The demand by optimizing wealthy individuals and large firms for whom the variable part of the brokerage costs are almost zero. For these, the transactions balances are determined by factors not in the Baumol model. The relevant factors could be the requirement for payments in money to individuals in category 1 or for transactions for which the

requirement is to pay in currency (for example, for bus fares), or minimum balances required by banks to keep a demand deposit account.

The electronic, regulatory and institutional innovations in recent years have blurred the distinction between demand deposits and various near-monies, and thereby shifted the transactions demand for the former. The invention and use of devices such as electronic or smart cards is reducing the need to hold notes and coins for small expenditures, thereby reducing the demand for currency.

The inventory demand function is the core implication of the Baumol model. It was derived under rather special and restrictive assumptions. As this chapter has shown, relaxing these assumptions tends to change the implied elasticities of demand. However, in general, the qualitative conclusions remain: in the aggregate, the demand for real transactions balances increases less than proportionately with real expenditures, decreases with the yield on alternative assets, and does not change if all prices change proportionately.

8 The demand function for money

A number of issues have to be resolved prior to the empirical estimation of the demand for money. Among these are the use and estimation of expected and permanent income and the treatment of lags in money demand. For the former, this chapter covers the use of rational expectations. Adaptive expectations are used for the measurement of permanent income. Costs of adjusting money balances lead to lags in the adjustment of actual to desired money balances. The simplest forms of lags are the first-order and second-order (linear) partial adjustment models.

This chapter also extends the money demand function to the open economy and investigates currency substitution and capital mobility.

Key concepts introduced in this chapter

- Permanent income
- Expected income
- Rational expectations
- Adaptive expectations
- General autoregressive model
- Lucas supply rule
- Keynesian supply function
- Partial adjustment models
- Autoregressive distributive lag model
- Currency substitution and capital mobility

Milton Friedman’s money demand function, presented in Chapter 2, argued that permanent income is one of the determinants of the demand for money. Other studies assume that the individual’s planned money balances are a function of his expected income during the period ahead. While the data on the actual past and present levels of national income is readily available, data on the expected and permanent income are not observable. This data has to be either generated or proxied in estimating the demand function for money.

Further, while the theoretical analyses of Chapters 2 to 6 provided the three basic specifications of the demand function for desired balances, there could be significant costs of reaching the desired levels in each period, so that the actual balances held may differ from those desired. This leads to the consideration of partial adjustment and lags in the money-demand function. Since our aim is to explain the actual balances held, the differences between the desired and actual money holdings and the procedures for handling the lags that occur in this process need to be examined.[1] In recent years, these issues have been pushed aside, though not addressed, by the increasing use of cointegration and error-correction estimation techniques.

While the money demand analyses of the preceding chapters established the arguments of the money demand function, they did not specify its specific functional form. This chapter introduces three of its more commonly used basic functional forms in empirical analyses for the closed economy. It then proceeds to the money demand function for the open economy under the heading of currency substitution.

Section 8.1 starts with three basic money demand functions, with actual income, expected income and permanent income as the scale variable. For the latter, Section 8.2 presents the rational expectations hypothesis for estimating expected income. Section 8.3 presents the adaptive expectations procedure for deriving permanent income and Section 8.4 lists the regressive and extrapolative procedures. Sections 8.5 to 8.8 present the partial adjustment model and the general autoregressive model. Section 8.9 focuses on the money demand function in the open economy.

Basic functional forms of the closed-economy money demand function

Monetary theory provides the variables that determine money demand but does not specify the particular form of the money demand function. The analysis of the demand for money in the preceding chapters implied that this demand depends on an income or wealth variable, often also called the “scale variable,” and on the rates of return on alternative assets. Since these rates of return are closely related to each other, so that including several of them in the same regression induces multicollinearity (discussed in the next chapter), the money demand equation that is usually estimated avoids multicollinearity by simplifying the estimating equation to include only one interest rate. With this simplification, and using actual income as the simplest form of the scale variable, the money demand function is:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(\,y,\,R)$$

where:

m^{d} = demand for real balances
y = actual real income
R = nominal interest rate

There is no real theoretical basis for assuming the form of this function to be linear, log-linear or non-linear in some other way. However, for reasons of convenience in estimation, the linear and log-linear functional forms are the most commonly used ones. This section compares these functional forms and points out the differences between them. It ignores, for simplification, the possibility of lags and expectations and assumes that money demand depends only upon current income and a nominal interest rate.

To start, consider the following simple specific forms of the money demand function, with μ as the random term. The subscript *t* has been omitted as being unnecessary for the discussion.

$$M/Y = a_0 + a_R R + \mu \tag{1}$$

$$M = a + a_R R + a_y y + a_P P + \mu \tag{2}$$

$$m = a + a_R R + a_y y + \mu \tag{3}$$

(1) assumes that the elasticity of the demand for money with respect to nominal income – and hence with respect to both prices and real income – is unity. (3) assumes that this elasticity is unity with respect to the price level but not necessarily so with respect to real income. (2) does not make either assumption.

(3) is the only function consistent with the discussion in earlier chapters that the individual's demand for money balances is in real rather than in nominal terms. Proceeding further with (3), money demand in a world where commodities and money are substitutes would also depend upon the expected rate of inflation π^e, so that (3) would be modified to:

$$m = a_0 + a_R R + a_y y + a_\pi \pi^e + \mu \tag{4}$$

Other variables, such as the expected exchange rate depreciation to take account of currency substitution in the open economy, as is done later in this chapter, could be introduced in a similar manner on the right-hand side of (4).

The money-demand functions are often estimated in a log-linear form. The log-linear form corresponding to (3) would be:

$$\ln m = \ln a_0 + \alpha \ln R + \beta \ln y + \ln \mu \tag{5}$$

A variant of (5) replaces ln *R* by ln (1 *R*) since *R* is usually between 0 and 1, so that its logarithmic value would be a negative n+umber whereas ln (1 *R*) would be positive. (5) is identical to:

+

$$m = a_0 R^\alpha y^\beta \mu \tag{6}$$

This functional form is the well-known *Cobb–Douglas* functional form. It was implied by the inventory analysis of the transactions demand for money, though not by the speculative or the precautionary demand analyses. In (5) and (6), the elasticity of the demand for real balances is α with respect to *R* and β with respect to *y*. A variant of (6) is:

$$\ln m = \ln a_0 + \alpha R + \beta \ln y + \ln \mu \tag{7}$$

(7) does not require taking the log of the interest rate since doing so would yield negative values when the values of *R* lie between 0 and 1. However, note that (7) translates to:

$$m = a_0 e^{\alpha R} y^\beta \mu \tag{8}$$

Since (6) and (8) are different and are unlikely to perform equally well, the researcher has to choose between them. There is no theoretical basis for doing so, with the result that the one that gets to be reported often depends upon its relative empirical performance for the data being used.

Scale variable in the money demand function

Current income as the scale variable

The linear form of the demand function for real balances, with current income as the scale variable, is:

$$m^{d}_{t} = a_0 + a_y y_t + a_R R_t + \mu_t \qquad a_0, a_y > 0, a_R < 0 \tag{9}$$

where μ is the random disturbance. (9) would become log-linear if each of the variables and μ_t were in logs.

Expected income as the scale variable

Another money demand function that is in common usage replaces current income by expected income. A demand function with expected income as its scale argument is:

$$m^{d}_{t} = a_0 + a_y y^{e}_{t} + a_R R_t + \mu_t \qquad a_0, a_y > 0, a_R < 0 \tag{10}$$

In (10), at the *beginning* of the period, m^{d}_{t} are the planned real balances for the period ahead, y^{e}_{t} is the expected income for the period. While we could have also introduced the interest rate in terms of its expected value, this is rarely done. The current practice is to estimate expected income y^{e}_{t}using the rational expectations hypothesis (REH).

Permanent income as the scale variable

As against current or expected income, Friedman's (1956) theoretical analysis of the demand for money presented in Chapter 2 implied that this demand depends upon wealth, or its proxy, permanent income, and on interest rates. For Friedman's analysis, the basic form of the demand function for real balances with permanent income is:

$$m^{d}_{t} = m^{d}(y^{p}_{t}, R_t)$$

where y^{p}_{t} is permanent income, which can be interpreted as the average expected income over the future. The simplified linear (or log-linear) form of this demand function for real balances is:

$$m^{d}_{t} = a_0 + a_y y^{p}_{t} + a_R R_t + \mu_t \qquad a_0, a_y > 0, a_r < 0 \tag{11}$$

Since data on the observed values of y^{p} does not normally exist, Friedman used the adaptive expectations hypothesis for deriving permanent income. Though the REH can be used as an alternative procedure for doing so, adaptive expectations seem more appropriate for estimating permanent income since the latter is best interpreted as the *average* expected value of income, rather than merely as expected income for the period ahead. Correspondingly, m^d_t in (11) should be interpreted as the *average* expected amount of desired real balances. The adaptive expectations procedure for constructing permanent income is explained in Section 8.3.

Note that the three scale variables in (9) to (11) are different, so that their estimation will yield different coefficients. Further, even their stability properties may differ. As discussed later in this chapter and in Chapter 9, the time series for several variables, including money and income, tend not to be stationary. The appropriate technique for such variables is cointegration analysis, which is a maximum likelihood vector autoregressive (VAR) technique. Such estimation ignores altogether the distinction between expected and permanent income, so that the prior application of the rational and adaptive expectations procedures is not needed if the money demand function is estimated using cointegration techniques. However, its accompanying error-correction estimation does incorporate adjustment lags and adaptive expectations.

Rational expectations

Theory of rational expectations

The rational expectations hypothesis (REH), first proposed by Muth (1961), is stated in various forms. One way of stating it is that the individual uses all the available information at *his* disposal in forming his expectations on the future values of a variable. Since individuals often have to – or choose to – operate with very limited information, the relevant information set is sometimes specified to be one of maximizing profit. In any case, the available information set is assumed to include the knowledge of the *relevant theory*,[2] with the rationally expected value of the variable being its value as predicted by this theory. The REH asserts that deviations of the actual from the theoretically predicted value will be randomly distributed with a zero mean and be uncorrelated with the available information and with the theoretically predicted value.

Note that the relevant theory will commonly determine the non-random prediction of a variable as a function of the parameters, the past values of the endogenous variables and the past, current and future values of the exogenous variables. Of these, the future values of the exogenous variables will usually not be known to the individual and their rational expectations values will be needed, so that the relevant theory for them will also have to be specified. In practical terms, the REH can be restated as: the expected values of the endogenous variables will be those predicted by the relevant theory, given the data on the past values of the endogenous variables, those on the past and current values of the relevant exogenous variables, and the rationally expected future values of the relevant exogenous variables.

Designate the rationally expected value of y^{e}_t predicted by the relevant theory as y^{T}_t, where the superscript T stands for the relevant theory. Since y^{T}_t takes account of all the information available to the individual, the REH asserts that the deviation of the actual value y_t from y^{T}_t will be random with a zero expected value and will be uncorrelated with the available information and, therefore, with y^{T}_t which is based on that information. The following incorporates the above statements in a set of simple equations to show the various assumptions and steps in deriving the rationally expected value y^{e*}_t of the variable y_t .

Since the rationally expected value y^{e*}_t – with ^{e*} standing for the rationally expected value – is assumed to be determined by the value y^{T}_t predicted by the relevant theory T, we have:

$$y^{e*}_t = y^{\mathrm{T}}_t \tag{12}$$

Since the REH assumes that the actual value y_t differs from the prediction of the relevant theory T by an error that is random and not correlated with any available information, we have:

$$y_t = y^{\mathrm{T}}_t + \eta_t \tag{13}$$

where:

$$E\eta_t = 0 \tag{14}$$

$$\rho(y^{\mathrm{T}}_t, \eta_t) = 0 \tag{15}$$

y_t = actual income
y^e_t = expected income
y^{e*}_t = rationally expected value of income
y^{T}_t = expected income predicted by the relevant theory
$E\eta_t$ = mathematical expectation of η_t
$\rho(y^{\mathrm{T}}_t, \eta_t)$ = correlation coefficient between y^{T}_t and η_t

(12) and (13) imply that:

$$y_t = y^{e*}_t + \eta_t \tag{16}$$

Taking the mathematical expectation of (16), with $E\eta_t = 0$ from (14), and using (12) gives:

$$Ey^{e*}_t = Ey_t = y^{\mathrm{T}}_t \tag{17}$$

If y^{e*}_t and y^{T} are assumed, as is often done, to be single valued, (17) becomes:

$$y^{e*}_t = Ey_t = y^{\mathrm{T}}_t \qquad (18)[3]$$

To implement (18) empirically, the rationally expected value y^{e*}_t can be obtained by estimating y_t using the function implied by the theory for its determination, and taking its expected value Ey_t.[4] This procedure is illustrated below and will also be applied in Chapter 17 in a macroeconomics context.

The "relevant theory"

A fundamental question in applying the REH is about the definition of the term "relevant theory." To an economist who believes that the economy tends to be at full employment, even though it is currently not in that state, the relevant theory for forming expectations on aggregate output is that the economy will be at the full-employment level. Consequently, the full-employment output will be the rationally expected one, so that the appropriate procedure would be to solve the model or theory for its full-employment state and substitute it for the expected output or real income. This procedure is the one adopted by economists in the modern classical approach.

However, for economists who believe that the economy is rarely, if ever, exactly in full employment, the rational expectation of next period's real income will not be one of full employment. The theory needed for their rational expectations of income would be a theory of the non-random part of the expected level of actual income, since this is the level that would differ from next period's actual income by a random term. Keynesian economists follow this line of thinking and need to specify a theory of the expected value of actual output for the period in question.

Hence, the application of the REH will yield different values of rationally expected output, depending upon the underlying assumption of the continuous existence or frequent absence of full employment. While the REH at the conceptual level can be and is used by both classical and Keynesian economics, its application, even in the context of an otherwise identical model (e.g. the IS–LM one), provides different predictions of the expected future income for the two approaches.

To proceed further, (18) can be used to construct the estimate of $y^{e*}{}_t$ by using the relevant theory to specify the determination of $y^{T}{}_t$. We illustrate this use of the theory by incorporating two alternative theories on the relationship between output and the rate of increase in the money supply. The first theory will be the Lucas supply rule, which underpins modern classical macroeconomics, and the second one will be the Keynesian theory.[5]

Information requirements of rational expectations: an aside

There is considerable dispute in the literature about the information requirements for rational expectations. The information available to any given individual varies considerably, *inter alia*, with the individual's level of education and interest, the openness of the society and the operating technology of information, as well as the losses from basing actions on inadequate, vague and inaccurate information. The actual amount of information at the disposal of the individual can vary from almost non-existent hard information[6] to extensive knowledge. The REH is meant to apply to all cases, regardless of the extent and accuracy of the available information.

Skeptics about the REH have argued that it requires that the possible future outcomes are well anticipated and that economic agents are assumed to be superior economists and statisticians, capable of analyzing the future general equilibrium of the economy (Arrow, 1978). However, the supporters of rational expectations reject such criticism and claim that:

> The implication that economic agents or economists are omniscient cannot fairly be drawn from Muth's profound insights. … Rational expectations are profit maximizing expectations. … If the past proves to be a very imperfect guide to the future, then theory and practice will be inaccurate.
>
> (Kantor, 1979, p. 1424).

> It is, however, incorrect to assume that rational expectations regards errors as insignificant or absent. The implication of rational expectations is that the forecast errors are not correlated with anything that could profitably be known when the forecast is made.
>
> (Kantor, 1979, p.1432).

Another view of rational expectations is provided by Robert Lucas, who popularized its usage in macroeconomics. The *Economist* website reported that Lucas at one time said that:

> [Rational expectations] doesn't describe the actual process people use trying to figure out the future. Our behavior is adaptive. We try some mode of behavior, if it is successful, we do it again. If not, we try something else. Rational expectations describe the situation when you've got it right.[7]

This interpretation means that, for most of the time spent in figuring out the future and acting on one's expectations, we would not have rational expectations with its critical property that the errors between the actual and the expected value would be random. Since rational expectations will only hold eventually ("when you have got it right"), they should be restricted only to the long-run analysis. They will not apply over short periods and in the short run. This interpretation of rational expectations is not consistent with the macroeconomic literature on it, including Lucas's own contributions on the short-run macroeconomic model. It is also not consistent with the use and analyses of the Lucas supply rule presented in this chapter and in Chapter 14. We shall henceforth ignore this interpretation.

Assessing the validity of the rational expectations hypothesis

The insight behind rational expectations at its conceptual level – that is, when an individual's expectations are based on all his available information – is undeniable. However, part of the available information comes from our understanding of the past and the present, which is itself incomplete and imperfect, as witnessed by the prevalence, even in hindsight, of different theories to explain any given observation. In addition, knowledge of the future is even more uncertain; as the quote at the end of this subsection argues, for the future, we don't even know what we don't know. Given the increasing increment of this degree of ignorance for periods further ahead, short-term (i.e. for the next quarter or so) predictions tend to do better than for periods further ahead. But, for these, the persistence forecast (i.e. the immediate future will be like the immediate past except for random variations) does quite well – and usually better than predictions based on any theory.

In the rational expectations hypothesis, the leap from the "subjective/personal theory" based on the available information to the assumption of the "relevant theory," common to all individuals as well as being the accurate theory, is a massive one. This assumption is also likely to be invalid. The exponents of the REH, with Kantor among them, focus on the former, whereas its critics, with Arrow among them, focus on the latter. Leaving aside the doctrinal disputes, the *empirical issue* boils down to a question of the usefulness or profitability of *acting* on one's rational expectations. This usefulness can be extremely limited when – without knowledge of the relevant theory and without good reliable information on the past values of the endogenous and exogenous variables, or on the relevant future values of the exogenous variables – the known paucity of information indicates that the actual error in the rationally expected value of a variable can be large[8] relative to the mean expected value of the variable, so that acting on the basis of the rationally expected value of the variable may not turn out to be a prudent exercise.[9] Conversely, if the information available is quite complete and the subjective probabilities are known to approximate the objective ones, the rational expectations could be an appropriate basis for action.

An interesting take on the nature of uncertainty and how it limits the reliability and usefulness of rational expectations for decisions is provided by the following quote:

> There are *things that we know*. There are [also] *known unknowns*; that is to say that there are things we now know that we don't know. But there are also *unknown unknowns* – things that we do not know we don't know. So when we do the best we can and we pull all the information together, and we then say, "Well, that is basically what we see as the situation," that is really only the known knowns and the known unknowns. And [as time passes] we discover a few more of those unknown unknowns. There is another way to phrase that, and that is that the absence of evidence is not evidence of absence.
>
> (Donald Rumsfeld in a news conference, June 2002; italics added).

Using the REH and the Lucas supply rule for predicting expected income

This rule assumes the modern classical model, with the labor market being in long-run equilibrium at full employment and with deviations in real national output from its full employment level y^f occurring only due to errors in predicting the actual level of the money supply. One form of the Lucas supply rule[10] specifies the relevant theory for the determination of output *y* in period *t* as:

$$y^{T}_{t} = y^{f}_{t} + \gamma(M_{t} - M^{e}_{t}) \tag{19}$$

where:

y^{f}_{t}= full employment level of output in *t*
M_t = nominal money stock in *t*
M^{e}_{t}= expected value of the nominal money stock in *t*
M^{e*}_{t} = rational expectations of M_t , formed in $t-1$

so that the rational expectation of income, with the Lucas supply rule as the relevant theory for its determination, is:

$$y^{e*}_{t} = y^{f}_{t} + \gamma(M_t - M^{e*}_{t}) \tag{20}$$

Use of (20) for predicting rationally expected income requires using the relevant theory to determine M^{e*}_{t}. The relevant theory depends upon the monetary policy being pursued by the monetary authority.[11] In the context of an exogenous money supply, the central bank controls the money supply and can determine the money supply in the economy on the basis of a "rule" or function. Assume this to be the case, and that the relevant theory for the central bank's money supply rule M^{T} is:

$$M^{T}_{t} = \Psi_0 + \Psi_1 u_{t-} + \Psi_2 M_{t-} \tag{21}$$

where u_t is the unemployment rate (or the output gap between full employment and actual output) in period *t*. Designating the random error in M_t as ξ_t, (21) leads to the specification of M_t as:

$$M_t = \Psi_0 + \Psi_1 u_{t-1} + \Psi_2 M_{t-1} + \xi_t \tag{22}$$

Estimating (22) will provide the estimated values of the coefficients Ψ_i, *i* 0, 1, 2, as $\hat{\Psi}$. These estimated coefficients can be used to estimate EM_t, which yields the rationally expected value M^{e*}_{t}, as:

$$\hat{M}^{e*}_{t} = E\hat{M}_t$$

$$= \hat{\Psi}_0 + \hat{\Psi}_1 u_{t-1} + \hat{\Psi}_2 M_{t-1} \tag{23}$$

Since $M_t - M^{e*}_{t} = M_t - \hat{\Psi}_0 + \hat{\Psi}_1 u_{t\ 1} + \hat{\Psi}_2 M_{t\ 1} = \hat{\xi}_t$, where $\hat{\xi}_t$ is the estimated value of ξ_t, (19) implies that:

$$y_t = y^{f}_{t} + \gamma\hat{\xi}_t + \eta_t \tag{24}$$[12]

In the estimation of (24), y^{f}_t is replaced by a constant term. The estimation of (24) then yields the estimated values $\hat{y}^{\mathrm{f}}_t$ and $\hat{\gamma}$, so that the rationally estimated value $y^{\mathrm{e}*}_t$ of y^{e}_t can now be derived from:

$$\hat{y}^{\mathrm{e}*}_t = \hat{y}^{\mathrm{f}}_t + \hat{\gamma}\,\hat{\xi}_t \qquad (25)^{13}$$

The procedure for the estimation of the money demand function using the REH and the Lucas supply function

In the above illustration of the REH, it was necessary to estimate the money supply function (22) in order to estimate the error in the expected value of the unanticipated money supply; then to use this value in (24) to estimate the expected value of real output/income; this was followed by the use of this estimated value of real income in the regression for the money demand function (10). Hence, estimating the money-demand function – using the REH and the Lucas supply rule – required estimation of at least three equations in a stepwise procedure. The reliability of its estimates of the money demand coefficients in (10) would therefore depend upon the proper specification of the model for y^{T}_t and of its subsidiary equations for the money supply function, as well as of the reliability of the data and the estimating techniques used at the various stages. Clearly, there is considerable scope for possible errors in specification and biased estimation.

Keynesians believe that one of the sources of errors in the above estimation is the specification of the Lucas supply rule as the "relevant theory" for the determination of income. They believe that a Keynesian supply function is the appropriate theory. The following presents their approach.

Using the REH and a Keynesian supply function for predicting expected income

A simple form of the Keynesian supply rule[14] for the context of an exogenous money supply is:

$$y^{\mathrm{T}}_t = y_{t-} + \beta(Dy_{t-}) \cdot M_t \qquad \beta \geq 0 \qquad (26)$$

where D$y_{t\ 1}$ (y^{f} $y_{t\ 1}$). (26) specifies that real income/output depends upon the actual money supply, rather than only on the unanticipated change in the money supply. Further, this impact depends on the prior state of the economy, with this state captured by the lagged output gap[15] D$y_{t\ 1}$. If the prior state is one of full employment, D$y_{t\ 1}$ 0, so that changes in the money supply will not change output. The larger the output gap, the larger the impact of the money supply on output.

The stochastic form of this relationship is:

$$y_t = y_{t-1} + \beta(Dy_{t-1})M_t + \eta_t \tag{27}$$

In estimation, (27) uses the *actual* value of M_t as a regressor and therefore does not require the prior estimation of the coefficients of the money supply function or of the anticipated money supply.[16] Consequently, (27) requires the estimation of only one equation for estimating the expected value of income, rather than estimation of two under the Lucas supply rule. Again assuming the REH, using the estimated value of β from (27) provides the estimate of the (Keynesian) rationally expected value y^e_t as:

$$\hat{y}^{e*}{}_t = \hat{y}_{t-1} + \hat{\beta} M_t \tag{28}$$

Compare (25) and (28). Note that both provide the "rationally expected values of income" but under different theories as being the "relevant" one.

The procedure for the estimation of the money demand function using the REH and the Keynesian supply function

Proceeding further with (28), the rationally expected value of y_t can now be inserted into the money demand function (10) to estimate the latter. Hence, the use of the Keynesian supply function and the REH requires only a two-step procedure for estimation of moneydemand.

Rational expectations – problems and approximations

While rational expectations require that y^e_t be based on all available information, the information available to the economist is different from that available to the individual. Further, the economist generally deals with aggregates – for example, with aggregate money demand or national income – rather than with the money demand or income of any given individual, so that what the relevant information set should be is not always clear. Furthermore, there are disputes among economists as to the relevant theory, or at least to the theory held by the public.[17] Even when there is agreement among economists – admittedly a rare occurrence – on the general form of the theory, there is usually disagreement on the values of the coefficients of the model *and* on the *expected* values of the exogenous variables for the period ahead. Even the data on the lagged values of the endogenous variables is usually approximate and subject to revision, sometimes for several years after the data period. These problems and disputes render rational expectations a blunt procedure at the estimation level, and its applications subject to doubt and disputes.

In view of the absence of direct quantitative data on expected income and problems with applying rational expectations at the empirical level, some researchers choose to proxy y^e_t in various ways. Two examples of this are:

1 Use the actual income y_t as a proxy for y^e_t, since the two differ only by a random term whose expected value is zero under rational expectations.

2 Use the autoregressive model:

$$y_t = \delta_0 + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \cdots + \mu_t \tag{29}$$

and then use $y^e_t = Ey_t$ and the estimated coefficients of (29) to estimate y^e_t. The justification for this approximation to rational expectations is that the past experience of income itself is likely to be the dominant part of the relevant information set of the individual and the public, and the past values of income are likely to be most important determinant of current income in the relevant model.

While the REH at the conceptual level is very appealing, such approximations in empirical applications do reduce its distinctiveness from the rivals to the REH and are not recommended – unless there is no better choice.

Adaptive expectations for the derivation of permanent income and estimation of money demand

The specification of permanent income

In order to illustrate the application of adaptive expectations in money demand estimation, we shall use permanent income as the income variable in the money demand function. This function is:

$$m^d_t = a_0 + a_y y^p_t + a_r R_t + \mu_t \qquad a_0,\ a_y > 0,\ a_r < 0 \tag{11}$$

The general adaptive expectations model assumes that the individual bases his permanent income on his experience of current and past actual income, so that the general function for permanent income y^p_t would be:

$$y^p_t = f(y_t, y_{t-1}, y_{t-2}, \ldots) \tag{30}$$

A simple form of (30), which has proved to be convenient for manipulation and was used by Friedman for deriving permanent income in his empirical work on consumption and money demand, is the *adaptive expectations (geometric distributed lag) function*. It specifies the functional form of y^p_t as:

$$y^p_t = \theta y_t + \theta(1-\theta)y_{t-1} + \theta(1-\theta)^2 y_{t-2} + \cdots \tag{31}$$

where $0 \le \theta \le 1$. Permanent income is thus specified as a weighted average of current and past incomes, with higher weights attached to the more recent incomes. Note that if $\theta = 0.40$, a weight often cited as approximating reality for annual consumption data, the weights decline in the pattern 0.4, 0.24, 0.144, 0.0864, …, so that income more than four years earlier can be effectively ignored. If actual income becomes constant, permanent income will come to equal this constant level of actual income.

Koyck transformation of the geometric distributed lag function

Lag (31) one period and multiply each term in it by $(1 - \theta)$. This gives:

$$(1-\theta)y^{p}_{t-1} = \theta(1-\theta)y_{t-1} + \theta(1-\theta)^2 y_{t-2} + \theta(1-\theta)^3 y_{t-3} + \cdots \quad (32)$$

Subtracting (32) from (31) gives the equation:

$$y^{p}_{t} = \theta y_t + (1-\theta)y^{p}_{t-1} \quad (33)$$

(33) is known as the *Koyck transformation*. This transformation allows permanent income to be stated in terms of the revision of its value last period in the light of current income.

Deriving the estimation form of the money demand function

Substituting y^{p}_{t} from (33) in the money demand function (11) gives:

$$m^{d}_{t} = a_0 + a_y\theta y_t + a_y(1-\theta)y^{p}_{t-1} + a_R R_t + \mu_t \quad (34)$$

Lag each term in (34) by one period and multiply it by $(1 - \theta)$. This gives:

$$(1-\theta)m^{d}_{t-1} = (1-\theta)a_0 + a_y(1-\theta)y^{p}_{t-1} + a_R(1-\theta)R_{t-1} + (1-\theta)\mu_{t-1} \quad (35)$$

Subtracting (35) from (34) to eliminate y^{p}_{t-1} gives:

$$m^{d}_{t} = a_0\theta + a_y\theta y_t + a_R R_t - a_R(1-\theta)R_{t-1} + (1-\theta)m^{d}_{t-1} + \{\mu_t - (1-\theta)\mu_{t-1}\} \quad (36)$$

where $a_y, a_R > 0$, and $0 \quad \theta \quad 1$. The objective in carrying out the above steps was to eliminate the variable y^{p} on which data is not available. (36) achieves this objective.

The estimating form of (36) is:

$$m^{d}_{t} = \alpha_0 + \alpha_1 y_t + \alpha_2 R_t + \alpha_3 R_{t-1} + \alpha_4 m^{d}_{t-1} + \eta_t \quad (37)$$

where:

$\alpha_0 = a_0\theta$
$\alpha_1 = a_y\theta$
$\alpha_2 = a_R$

$\alpha_3 = -a_R(1 - \theta)$
$\alpha_4 = (1 - \theta)$

$\eta_t = \{\mu_t - (1 - \theta)\mu_{t-1}\}$

Note that (37) involves lagged terms in m and in R, but not in y. Further, the disturbance term in (37) is $\{\mu_t + (1-\theta)\mu_{t-1}\}$, which is a *moving average error*.

Adaptive expectations as the error-learning model

The adaptive expectations procedure in the form given by (33) can also be stated in a form known as the *error-learning model*. This form is:

$$(y^{p}_{t} - y^{p}_{t-1}) = \theta(y - y^{p}) \qquad (38)$$

which specifies the *revision in permanent income* on the basis of the experienced difference or "error" between the actual income in *t* and the permanent income for period (*t*_ 1). From (38), if $\theta = 0$, the estimate of permanent income is never revised on the basis of the experience of current income.

Assessing the relevance and validity of the adaptive expectations procedure

If we compare the rational and the adaptive expectations procedures for estimating the money-demand function, the former requires the estimation of at least two (possibly three, as in our illustration above) equations for the Lucas supply rule. However, doing so has the advantage that it allows better identification of the sources of shifts. Conversely, the adaptive expectations procedure has the disadvantage that if the parameters of the estimated money demand function shift, it is not clear whether the parameters of the money demand function or of the expected income equation have shifted. Further, in cases of monotonically increasing (decreasing) income paths, adaptive expectations induce persistent and increasing negative (positive) errors (i.e. $y_t \quad y^{\mathrm{p}}_t$) in expected income relative to actual income, so that rational individuals will revise their procedure for forming expectations away from adaptive expectations. Adaptive expectations also fail to take account of any information available to the individual about future changes in income, and are said to be (only) *backward looking*.

However, note that the adaptive expectations model, in spite of its name, really provides an estimate of the average level of future income – rather than the expected value of income for the period ahead – through its geometric distributed lag procedure, while the REH procedure provides a more appropriate estimate of the latter. The two procedures therefore provide proxies for different concepts of income, so that the choice among them should depend on the income variable, which is the appropriate scale variable in the money demand function. If the non-stochastic component of income is fluctuating and the appropriate scale variable is permanent or *average expected* income y^{p}_t, the geometric distributed lag would be a better representation of this average than the rationally expected value of current income $y^{\mathrm{e}*}_t$.[18]

Regressive and extrapolative expectations

An alternative to adaptive expectations is the *regressive expectations* model, which specifies that:

$$y^{\mathrm{e}}_t = y_{t-} + \delta(y^{\mathrm{LR}} - y_{t-}) \tag{39}$$

where y^{LR} is the long-run level of income. Here, the expectation is that income will tend towards its long-run value.[19]

Another model of expectations is the *extrapolative expectations* one. It is that:

$$y^{\mathrm{e}}_t - y_{t-} = \delta(y^{t-} - y^{t-}) \tag{40}$$

This model assumes that income is expected to change as a proportion of the change in income in the preceding period. That is, recent *changes* – or the factors producing those changes – are expected to determine the pattern of future changes.

Whether the adaptive, regressive, extrapolative or rational expectations procedures are more appropriate depends upon how the individual forms his expectations. The adaptive expectations model seems to be the most common one for modeling permanent income, i.e. average expected income, while the rational expectations procedure is the most common one for modeling expected income over the period ahead.

Lags in adjustment and the costs of changing money balances

Lags often occur in the adjustment of money demand to its desired long-run value. There can be several reasons for such lags. Among these are: (i) habit persistence and inertia, (ii) slow adjustment of money balances due to uncertainty on whether the changes in the determinants (income and interest rates) of money demand are transitory or longer lasting, and (iii) adjustment costs, which can be monetary or non-monetary. This section focuses on adjustment costs and presents the derivation of adjustment patterns from adjustment cost functions.

First-order partial adjustment model

One reason for an adjustment lag can be the short-run cost of changing money balances. To investigate the relationship between such costs and the adjustment lag in money balances, let the individual's desired real balances be $m^*{}_t$ and assume that the individual faces various types of costs of adjusting instantaneously to $m^*{}_t$. Examples of such costs are:

(i) The cost of being below or above $m^*{}_t$. For example, having inadequate balances can prevent one from carrying out purchases which require immediate payments in money.

(ii) The cost of changing actual balances from m_{t-1} to m_t.

These costs can take various forms. A simple form of these occurs when (i) has the proportional quadratic form $a(m_t \ - m^*{}_t)^2$ and (ii) has the proportional quadratic form $b(m_t - m_{t-1})^2$. Assuming these to be so, the total adjustment cost c of reaching the desired balance in period t is given by:

$$c_t = a(m_t - m^*{}_t)^2 + b(m_t - m_{t-1})^2 \qquad a, b \geq 0 \tag{41}$$

The individual is taken to minimize this cost. The first-order condition for maximization is that:

$$\partial c_t / \partial m_t = 2a(m_t - m^*{}_t) + 2b(m_t - m_{t-1}) = 0 \tag{42}$$

which yields the actual balances m_t as:

$$m_t = \gamma m^*{}_t + (1 - \gamma) m_{t-1} \tag{43}$$

where $\gamma = a/(a + b)$. (43) can be restated in a more intuitive form as:

$$m_t - m_{t-1} = \gamma (m*_t - m_{t-1}) \quad 0 \leq \gamma \leq 1 \tag{44}$$

(43) and (44) constitute the *first-order* (i.e. with a one-period lag only) *partial adjustment model* (*PAM*): the adjustment of real balances in period *t* is partial, linear and involves a one-period lag. This model suffers from the disadvantage that if m^*_t has a positive or negative trend, the divergence of actual balances from the desired ones will increase over time. Individuals would find it profitable to avoid this by abandoning the first-order PAM and using some other adjustment mechanism.[20] Therefore, the first-order PAM is inappropriate when the desired or actual balances have a strong trend component.[21] A higher-level PAM would be more appropriate in such a case.[22]

Second-order partial adjustment model

Higher-order partial adjustment models result from more complicated specifications of the adjustment costs. The *second-order* (i.e. with a two-period) *partial adjustment model* is given by the adjustment cost function:

$$c_t = a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 + k(Om_t - Om_{t-1})^2 \qquad a,\ b, k > 0 \tag{45}$$

$$= a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 + k(m_t - 2m_{t-1} + m_{t-2})^2 \tag{46}$$

where $Om_t\ m_t\ m_{t\ 1}$, and $k(Om_t\ Om_{t\ 1})^2$ is additional to the adjustment costs (i) and

$= \quad - \quad =-$

(ii) and represents the cost of continual adjustments over time in balances. Minimizing (46) – that is, setting the partial derivative of *c* with respect to *m* equal to zero – and solving implies that:

$$m_t = \gamma_1 m^*_t + \gamma_2 m_{t-1} + (1 - \gamma_1 - \gamma_2)m_{t-2} \tag{47}$$

where:

$\gamma_1 = a/(a + b + k)$

$\gamma_2 = (b + 2k)/(a + b + k)$

Since (47) has a two-period lag, it provides the second-order partial adjustment model.

Error feedback model

A further elaboration of these models is obtained if, in addition to the earlier types of costs, the costs of continual adjustment were less when the actual changes Om_t were in the same direction as the desired changes Om^*_t. A specification of such a cost function would be:

$$c_t = a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 - kOm^*_t(m_t - m_{t-1}) \quad a, b, k > 0 \tag{48}$$

In this case, the demand for actual balances would be:

$$m_t = m_{t-1} + \gamma_1(m^*_t - m_{t-1}) + \gamma_2(m^*_t - m^*_{t-1}) \tag{49}$$

where γ_1 $a/(a$ $b)$ and γ_2 $k/$ $2(a$ $b)$. (49) is another form of PAM and is the *error feedback model.* $= [+]$ $= [\{ + \}]$

Assessing the validity of partial adjustment models

The various types of adjustment cost functions depend upon the notion that it is costly for the individual to change money balances. As in the inventory model of transactions balances in Chapter 4, this cost is the sum of both monetary and non-monetary costs and will differ for the different definitions of money. In practice, in modern financially developed economies with internet banking, the costs of converting to M1 from savings deposits and other near-money assets have become virtually zero, so that there should not be any significant adjustment lags at the level of the individual. The costs of changing M2 can be similarly very small and may be of little consequence at the individual's level. This is especially so when such costs are compared with those of changing the individual's stock of commodities or labor supplied.[23] The costs of adjustment for monetary aggregates usually become significant only when such adjustment involves converting bonds or commodities into a monetary asset. But this happens rarely for meeting desired changes in the demand for M1 and M2, so that the practice of using PAM models, especially for the narrower definition of money, may be of questionable value. However, the lagged adjustment implied by these models is part of the error-correction modeling within the currently popular cointegration technique.

Money demand with the first-order PAM

If there exist adjustment costs in changing money holdings, these costs should properly be incorporated into the structure of the individual's decision processes and the demand for money holdings be derived after such incorporation. However, this can prove to be analytically intractable, so that the usual procedure is to derive the demand function separately from the adjustment function and then combine them. This is the procedure followed here.

Assume that the individual's demand for real balances depends on current real income *y* and the nominal interest rate *R*, so that the demand function is:

$$m^*_t = a_0 + a_y y_t + a_R R_t + \mu_t \qquad a_0, a_y > 0, a_R < 0 \tag{50}$$

where μ is white noise. m^*_t are the desired real balances in the absence of adjustment

costs.

Further, assume the first-order PAM, which is:

$$m_t - m_{t-1} = \gamma(m^*_t - m_{t-1}) \quad 0 \le \gamma \le 1 \tag{44}$$

Substituting (44) into (50) to eliminate m^*_t gives:

$$m_t = a_0\gamma + a_y\gamma y_t + a_R\gamma R_t + (1 - \gamma)m_{t-1} + \gamma\mu_t \tag{51}$$

where a_0, $a_y > 0$, $a_R < 0$ and $0 \leq \gamma \leq 1$. The estimating form of (51) is:

$$m^d_t = \beta_0 + \beta_1 y_t + \beta_2 R_t + \beta_3 m_{t-} + \xi_t \tag{52}$$

where $\beta_0 = a_0\gamma$, $\beta_1 = a_y\gamma$, $\beta_2 = a_R\gamma$, $\beta_3 = (1 - \gamma)$ and $\xi_t = \gamma\mu_t$.

The estimating equations (37) and (52) should be compared to see the effects of adaptive expectations versus those of the first-order PAM on the estimated money demand equation. Each introduces the lagged money balances m_{t-1} into this equation, but adaptive expectations also introduce the lagged interest rate R_{t_1}. The disturbance terms also have different properties.

Money demand with the first-order PAM and adaptive expectations of permanent income

Assume now that money demand depends on permanent income, so that our model consists of the following three equations:

$$m^*_t = a_0 + a_y y^p{}_t + a_R R_t + \mu_t \qquad a_0, a_y > 0, a_R < 0 \tag{11}$$

$$y^p{}_t = \theta y_t + (1-\theta)y^p{}_{t-} \tag{33}$$

$$m_t = \gamma m^*_t + (1-\gamma)m_{t-1} \tag{43}$$

where (43) can be restated as:

$$m^*_t = (1/\gamma)m_t - \{(1-\gamma)/\gamma\}m_{t-1} \tag{53}$$

This model implies the estimating equation:[24]

$$m_t = a_0\theta\gamma + a_y\theta\gamma\, y_t + a_R\gamma\, R_t - a_R\gamma\,(1-\theta)R_{t-1} + (2-\gamma-\theta)m_{t-1}$$

$$-(1-\theta)(1-\gamma)m_{t-2} + \gamma\{\mu_t - (1-\theta)\mu_{t-1}\} \tag{54}$$

where a_0, $a_y > 0$, $a_R < 0$ and $0 \leq \gamma, \theta \leq 1$. The estimating form of (54) is:

$$m^d_t = \alpha_0 + \alpha_1 y_t + \alpha_2 R_t + \alpha_3 R_{t-1} + \alpha_4 m_{t-1} + \alpha_5 m_{t-2} + \eta_t \tag{55}$$

where:

$$\alpha_0 = a_0\theta\gamma$$
$$\alpha_1 = a_y\theta\gamma$$
$$\alpha_2 = a_R\gamma$$

$$\alpha_3 = -a_R\gamma(1-\theta)$$
$$\alpha_4 = (2-\gamma-\theta)$$

$$\alpha_5 = -(1-\theta)(1-\gamma)$$

$$\eta_t = \gamma \{\mu_t - (1 - \theta)\mu_{t-1}\}$$

(55) provides a more general estimating equation than either the PAM model or the adaptive expectations model. These two models are therefore nested in (55), with the PAM one alone obtained when $\theta = 1$ and the adaptive expectations obtained when $\gamma = 1$. Hence (55) provides a way of testing whether there exist both or either of these processes. However, (55) is not necessarily preferable to the alternative estimation procedure that uses the PAM and rational expectations to estimate expected income.

The structural coefficients in (54) are: a_0, a_y, a_R, γ, θ. The coefficients in the estimating equation (55) are: α_0, …, α_5. Hence, there are only five structural coefficients compared with six coefficients in the estimated equation, so that appropriate non-linear restrictions have to be imposed on the α_i in the estimating equation (55).

The earlier criticisms of adaptive expectations from the rational expectations perspective also apply here. To reiterate, the criticism is that adaptive expectations are backward looking and ignore information that may be available to the individual about the future as well as on other variables. Further, if the expectations parameter θ shifts, the estimating equation (55) would shift, without it being transparent whether the shift is due to a shift in γ, in θ, or in the coefficients of the demand function. By comparison, using the rational expectations procedures to estimate $y^p{}_t$ in the first step, and then estimating the demand function with PAM, will more clearly disclose the source of the shift in the money demand function.

Autoregressive distributed lag model: an introduction

Now suppose that the demand for real balances depends on the current and lagged values of real income and its own lagged values. That is, it has the form:

$$m_t = a_0 y_t + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + b_1 m_{t-1} + b_2 m_{t-2} + \cdots \tag{56}$$

This equation is the representation of the general *autoregressive distributed lag (ARDL) model*. $y_{t\ i}$ can be replaced by $L^i y_t$, where L^i is the lag operator, which can be treated as a variable subject to mathematical manipulation in the following manner:

$$\begin{aligned} a_0 y_t + a_1 y_{t-1} + a_2 y_{t-2} + \cdots &= a_0 y_t + a_1 L y_t + a_2 L^2 y_t + \cdots \\ &= y_t(a_0 + a_1 L + a_2 L^2 + \cdots) \\ &= a(L) y_t \end{aligned} \tag{57}$$

where $a(L)$ is the polynomial $(a_0 + a_1 L + a_2 L^2 + \cdots)$ in L. Hence, (56) can be rewritten as:

$$m_t = a(L) y_t + b(L) m_t \tag{58}$$

where:

$$a(L) = a_0 + a_1 L + a_2 L^2 + \cdots$$

$$b(L) = b_1 L + b_2 L^2 + \cdots$$

Therefore,

$$m_t - b(L) m_t = a(L) y_t$$

$$m_t = [\{1 - b(L)\}^{-1} \cdot a(L)]y_t \tag{59}$$

so that m_t becomes a function solely of y_t and its lagged terms, without its own lagged values, which are now omitted from the explanatory terms. (59) is the compact form of the ARDL lag model.

An illustration: a simple ARDL model

As an illustration, consider the simplest example of (59) where $a(L) = a_0$ and $b(L) = b_1L$. That is, (56) is simplified to:

$$m_t = a_y y_t + b_1 m_{t-1} \tag{60}$$

In this case, (59) simplifies to:

$$m_t = \{1 - b_1L\}^{-1} \cdot a_y y_t \tag{61}$$

Expand $\{1 \ b_1L\}^{-1}$ in a *Taylor's series* around $E(b_1L) = 0$, where $E(b_1L)$ is the mean value of b_1L. This gives:

$$\{1 - b_1L\}^{-1} = \{1 + b_1L + b_2L^2 + \cdots$$

Hence, (61) becomes:

$$m_t = \{1 + b_1L + b_2L^2 + \cdots\}a_y y_t \tag{62}$$

$$= a_y y_t + a_y b_1 y_{t-1} + a_y b_2 y_{t-2} + \cdots \tag{63}$$

While (60) and (63) are mathematical transformations of each other, so that their economic content is identical, the money demand function in the form of (63) does not contain the lagged value of the endogenous variable, although we started with equation (60) where it does so. Conversely, we could have started with (63) without the lagged money term and derived (60) as equivalent to it. Hence a comparison of (60) and (63) – and of (59) with (56) for the general case – leads to the caution that it may not be possible to distinguish between a money demand equation which contains the lagged values of the endogenous variable and other independent variables, and one which contains only the current and lagged values of the independent variables.

The general ARDL model with the suitable addition of disturbance terms is now in common usage in monetary analysis, and falls in the category of vector autoregression (VAR) models.[25] Its relationship with the now popular cointegration and error-correction estimation is given in the appendix to Chapter 9.

Demand for money in the open economy

This book has so far concentrated on the demand for money in the closed economy. This is the general pattern of studies on money demand. However, economies are becoming increasingly open to flows of commodities and financial assets, so that a special topic in the money demand literature deals with money demand in the open economy, in which economic units have access not only to domestic financial assets but also to foreign ones.

For portfolio investments in open economies, the financial alternatives to holding domestic money include the currencies and bonds of foreign countries, in addition to domestic bonds, so that the determinants of the domestic money demand should include not only the rates of return on domestic assets but also those on foreign assets. Since these assets include foreign money holdings, money demand studies for open economies need to pay special attention to substitution between domestic and foreign monies. This determination is especially relevant for open economies in which foreign currencies are extensively traded and foreign monies are part of the domestic media of payments. Note that the relevant literature on substitution between domestic and foreign money in the open economy uses the word "currency" for money. This chapter follows this usage.

Currency substitution (CS) can be defined as substitution between domestic and foreign currencies, which is "*currency–currency substitution.*" Substitution can also exist between domestic currency and foreign bonds, and between domestic currency and domestic bonds, which are "*currency–bond substitutions*." Designating, respectively, the nominal values of domestic money, foreign money, domestic bonds and foreign bonds by M, M^*, B and B^*, CS can be measured by $\partial M/\partial M^*$, while the various currency–bond substitutions would be measured by $\partial M/\partial B$, $\partial M/\partial B^*$, $\partial M^*/\partial B$ and $\partial M^*/\partial B^*$, or by their corresponding elasticities.

Giovannini and Turtleboom (1994), Mizen and Pentecost (1996) and Sriram (1999) provide extensive reviews of the CS literature.

Theories of currency substitution

The magnitude of CS will depend both on portfolio selection considerations – since both M and M^* are assets in a portfolio[26] – and on substitution between them as media of payments in the domestic economy. Therefore, the relevant approaches to the degree of CS are the portfolio/asset approach and the transactions approach.

For the asset/portfolio approach, the relevant theory would be the theory of portfolio selection, as set out in Chapter 5, which would treat M and M^* among the assets in the portfolio. This theory would determine substitution between currencies on the basis of their expected yield and risk. Two currencies would therefore be perfect substitutes if they had identical returns. They would be poor substitutes if, with identical risk, the return on one dominated that on the other. This identity of risk dominance does not in general apply in practice. Note that if some types of bonds were riskless, then, with a higher return, bonds would dominate over money, so that there would be zero portfolio demand for currency.

For the transactions approach to the demand for media of payments, it is the general acceptance in daily exchanges and payments that would determine the degree of substitution between the alternative assets. If the foreign currencies do serve as a medium of payments in the domestic economy, the classic demand analysis for the total of the media of payments, i.e. for the sum of M and M^*, is the Baumol–Tobin inventory analysis presented in Chapter 4. Under this approach, since domestic and foreign bonds do not serve as media of payments they would have a relatively low substitutability with the domestic currency, while that between M and M^* could be much higher. Further, the demand for $(M + M^*/\rho)$[27] would be a function of the domestic expenditures or GDP to be financed. For a given amount of transactions or expenditures to be financed, an increase in one medium of payments implies a decrease in the other, so that transactions demand analysis implies that $\partial M/\partial(M^*/\rho) < 0$. That is, in economies in which both M and M^* do act as media of payments, $\partial M/\partial(M^*/\rho)$ would be negative and significant. In the limit, if domestic residents are indifferent whether they receive payments in the domestic money or in the foreign one, $E_{M,M^*} = -1$, where $E_{M,M^*} = (M^*/M)(\partial M/\partial(M^*/\rho))$. This elasticity would be very much smaller in absolute magnitude, or non-existent, in open economies in which the usage of foreign currency for domestic payments involves significant additional costs to those for payments in the local currency. If this cost is sufficiently high, $E_{M,M^*} = 0$. Therefore, the magnitude of E_{M,M^*} is clearly likely to vary between economies which do not extensively use foreign monies in domestic payments for goods[28] and those economies in which the foreign money is extensively used as a medium of payments, alongside (or in preference to) the domestic money. "Partially dollarized economies" – defined as ones in which the domestic currency and the foreign one circulate side by side, with buyers and sellers indifferent between their use in settling transactions – are especially ones in which E_{M,M^*} tends to -1.[29]

Handa (1988) argued that economic agents in even very open economies but without effective dollarization tend to use the domestic currency as the preferred medium of payments and do not easily switch to the use of foreign currencies for payments because of the transactions costs[30] imposed on retail payments. He therefore designated the domestic currency as being the "preferred habitat"[31] for the domestic medium of payments. Under this hypothesis, the degree of substitutability between the domestic currency and a given foreign one would depend on the latter's acceptance for payments in the domestic economy or the cost and ease of conversion from the latter into the former. In general, there would be a very significant transactions cost in conversion of foreign currencies into the domestic currency. These costs lie in the spread between the purchase and sale conversion rates and in banks' commissions, and are usually quite significant for the size of the transactions of the representative household in the economy. Further, in retail transactions, payment in a foreign currency is usually at an unfavorable exchange rate set by the retailer. Consequently, the general presumption under the preferred habitat approach would be that foreign currencies will have low elasticities of substitution with the domestic currency, except possibly in special cases where a particular foreign currency is generally accepted in payments at par in the domestic economy. To illustrate, while sellers in Canada often accept US dollars, their offer by buyers is not all that common, because there is a greater cost to paying in the US dollar than is specified by the bank exchange conversion rate. Hence, under the transactions approach, the degree of substitution between the US dollar and the Canadian dollar need not be high and could be quite low.[32] The Canadian dollar finds almost no acceptance in the United States, so they are poor substitutes in the US economy. Further, in the Canadian economy, even if the Canadian and US dollars proved to be good substitutes, British currency is not generally accepted and would be a poor substitute for the Canadian dollar. Most open economies tend to be of this type, so that, except for special cases, the preferred habitat hypothesis implies that we should expect even quite open economies (open but without extensive usage of foreign currencies in domestic retail payments) to have $E_{M,M*}$ close to zero or with a small negative value.

Among the special cases of possibly high CS was the historical use of the local currency and the imperial one in colonies during the colonial era. Another special case is the use of the US dollar as a second medium of payments in domestic transactions in partially dollarized economies.[33] For such economies, the transactions demand for the media of payments implies that, for a given amount of transactions and GDP to be financed in economies in which both M and M * act as media of payments, a decrease in one would have to be offset by an equivalent increase in the other. Hence, partially dollarized economies are especially likely to have $E_{M\ ,M\ *}$ equal to –1, and an infinite elasticity of substitution,[34] while non-dollarized economies will have significantly lower elasticities of substitution.

Two broad approaches to CS: weak substitutability between monies and bonds

It is an implicit assumption of the CS literature that weak separability (see Chapter 7) exists between the four financial assets (domestic money, foreign money, domestic bonds and foreign bonds) and other goods, which include commodities and leisure, so that the demand functions for these four assets can be estimated by using only the returns on the four financial assets and the amount to be allocated among them. Proceeding further, the literature allows two possibilities:

A. Preferences over the domestic and foreign monies are not weakly separable from domestic and foreign bonds. That is, $U(M^*, M^*/\rho, B, B^*)$ is not weakly separable into a sub-function with M * and M */*ρ*. Estimations related to this hypothesis have been labeled in the CS literature the *"unrestricted approach."* As is discussed later, this approach is more suited to the portfolio approach than to the transactions one. In this approach, the demand function for domestic money will include the returns on all four assets, in addition to other variables, such as a scale variable.

B. Preferences over domestic and foreign monies are weakly separable from domestic and foreign bonds. That is, $U(M^*, M^*/\rho, B, B^*)$ is weakly separable into a sub-function with M^* and M^*/ρ, sothat:

$$U(M^*, M^*/\rho, B, B^*) = U(f(M, M^*/\rho), B, B^*).$$

Estimations related to this hypothesis have been labeled the "*restricted approach*" in the CS literature. This approach is appropriate for the demand for the two monies as domestic media of payments. It allows the possibility that domestic money and foreign money may act as media of payments in the domestic economy, but bonds do not.[35] If this is so, the demand functions for M and M^* can be estimated as a function of ρ, the returns on M and M^* and the amount to be allocated between them. Such estimation is said to be "restricted," since it is independent of the returns on bonds.

Estimation procedures and problems

There are three common methods of estimating currency substitution. These are:

- estimation of the elasticities of substitution
- estimation of a money demand function
- estimation of the ratio of domestic to foreign money balances.

Estimation of the elasticities of substitution

This procedure involves estimation of the Euler equations (first-order conditions) based on a constant (CES) or variable (VES) elasticity-of-substitution function. This method follows Chetty's procedure, explained in Chapter 7. In the unrestricted choice framework, the domestic money and foreign money balances, along with domestic and foreign bonds, would appear in the VES utility function. The estimating equations will be derived from the Euler conditions (see Chapter 7).[36] This procedure allows estimation of the elasticity of substitution between the two monies, and between the domestic money and the two types of bonds. The estimating equations in this case would be similar to those specified in Chapter 7 for the VES model, and are not listed here explicitly. In the restricted choice framework, with domestic and foreign monies weakly separable from bonds, the VES function would only include the two monies, so that the foreign money holdings of domestic residents would be regressed on domestic money balances and their "price" ratio. Among the studies based on this approach are those of Miles (1978) and Handa (1971).

Note that an alternative to the VES model is to assume a priori a unit elasticity of substitution between the assets and construct their chained Divisia or certainty-equivalence index (for which, see Chapter 7) with time-variant expenditure shares. These methods can be used for M and M^* only under the weak separability assumption of the restricted choice framework, or for all four financial assets for unrestricted choice. Estimation is not needed for the construction of the Divisia and certainty-equivalence aggregates.

Estimation of the domestic money-demand function

This estimation procedure is to expand the estimating money-demand equation to include among its regressors the return on at least one foreign currency, as well as returns on foreign bonds (and sometimes also physical) assets. This is the more common method of estimating currency substitution. It can be found in Bordo and Choudhri (1982), Bana and Handa (1987)[37] and Handa (1988).

For the unrestricted choice approach, the standard money-demand function, modified to take account of foreign currencies and foreign bonds as alternatives to domestic money, is usually specified as:

$$m^{\mathrm{d}} = \alpha_0 + \alpha_R R + \alpha_y y + \alpha_\varepsilon \varepsilon^{\mathrm{e}} + \alpha_{R*} R^* + \mu \qquad (64)$$[38]

where:

$m^{\mathrm{d}} =$ domestic money balances in real terms
$y =$ domestic real national income
$R =$ nominal yield on domestic bonds (= domestic rate of interest)
$R^{\mathrm{F}} =$ nominal interest rate on foreign bonds
$R^* =$ nominal yield on foreign bonds
(= foreign rate of interest + expected appreciation of the foreign currency)
$\rho =$ exchange rate (domestic currency per unit of foreign currency)
$\varepsilon^{\mathrm{e}} =$ expected return on the foreign currency
$\mu =$ disturbance term.

In (64), the three rates of return are ε, R and R^*. Note that the returns on domestic and foreign currencies include both their non-monetary returns – that is, their liquidity services, etc. – and the change in their nominal values relative to each other. While the liquidity and other non-monetary services are often critical for the demand for foreign currencies, data on them is usually non-existent, so that they are almost always excluded from the analysis. This is a significant deficiency of the empirical studies on currency substitution since, except in effectively dollarized economies, the acceptance in exchanges of domestic and foreign currencies and the ease of payment differ considerably.

Forced by the lack of data on the non-monetary/liquidity costs of domestic and foreign monies, the monetary return on foreign currencies is measured by the expected rate of appreciation of the foreign currency *vis-à-vis* the domestic currency.[39] This expected appreciation equals $(\_\partial \rho/\partial t)^{\mathrm{e}}$, where ρ is the number of units of the domestic currency per unit of the foreign one, so that $(\_\partial \rho/\partial t)^{\mathrm{e}}$ is the opportunity cost of holding the domestic currency rather than the foreign one. Therefore, ε^{e} in (64) is measured by $(\_\partial \rho/\partial t)^{\mathrm{e}}$. In empirical estimations using quarterly data, the proxy usually used for ε^{e} is $(\underline{F}\ S)/S$, where F is the 90-day forward exchange rate and S is the spot rate. For empirical studies on countries other than the USA, the foreign currency is usually taken to be the US dollar.

In open economies with perfect financial markets, the domestic and foreign interest rates are related by the interest rate parity (IRP) condition:

$$(1 + R_t) = (1 + R^{\mathrm{F}}{}_t)(1 + \varepsilon^{\mathrm{e}}{}_t) \tag{65}$$

where R^{F} is the rate of interest on foreign bonds and $\varepsilon^{\mathrm{e}}(= (\partial \rho/\partial t)^{\mathrm{e}})$ is the expected rate of depreciation of the domestic currency. The common approximation to (65) is:

$$R_t = R^{\mathrm{F}}_t + \varepsilon^{\mathrm{e}}_t \tag{66}$$

R_t, R^{f}_t and $\varepsilon^{\mathrm{e}}_t$ are all arguments of the open-economy money demand function. (66) implies that only two of these three variables are independent of each other, so that any two of them, but not all three, should be included in the estimating money demand equation. The two variables so selected are usually the domestic rate of interest and the expected exchange rate appreciation: many studies set $\alpha_{R*}=0$ prior to estimation, using the intuition that substitution between domestic money and foreign bonds is likely to be minimal.

(64) is usually estimated in a log-linear form, so that its coefficients represent elasticities. The cross-price elasticity α_R is the indicator of price-substitution[40] between the domestic currency and domestic bonds, and the cross-price elasticity α_{R*} is the indicator of price-substitution between the domestic currency and foreign bonds.[41] α_ε is the cross-price elasticity[42] with respect to the return on foreign currencies and can therefore be used as an indicator of CS.

In a four-asset portfolio selection analysis, the signs of any of the three cross-elasticities α_ε, α_R and $\alpha_{\rho*}$ are not specified by theory and could be negative or positive.[43] An empirically determined negative sign is interpreted as evidence of substitution between the domestic currency and the relevant asset. While a positive sign is sometimes interpreted as evidence of complementarity, this interpretation is not necessarily correct since it can reflect some other effect. This possibility is especially likely for the sign of α_R , as we illustrate later through the discussion on the substitution between M and M^* in the medium-of-payments role.

Estimation of capital mobility

Capital mobility is distinct from CS and may be defined as the net outflow of funds from the domestic economy into foreign assets, so that it would be specified by the overall substitution between the sum of the domestic currency and domestic bonds into foreign currency and foreign bonds. This would require the estimation of both the domestic money and bond equations. Therefore, the coefficients in the money demand function alone cannot be used as an indicator of capital mobility.[44]

The special relation between M and M in the medium-of-payments function*

Since the domestic currency and foreign bonds are both assets, portfolio selection theory implies that an increase in the return on foreign bonds relative to the return on the domestic currency (the riskless asset with a zero return) would cause substitution between them, thereby implying that $\partial M/\partial R^* = \alpha_{R^*} \leq 0$. This effect can be decomposed into two components specified by:

$$\frac{\partial M^d}{\partial R^*} = \frac{\partial (M^d)}{\partial R^*}\bigg|_{M^*=M^*} + \frac{\partial M^d}{\partial (M^*/p)} \frac{\partial (M^*/p)}{\partial R^*} \tag{67}$$

In (67), the first term on the right represents direct substitution between M and foreign bonds, holding foreign money balances constant. This (direct) effect occurs because the increase in foreign bonds increases the opportunity cost of holding domestic money relative to foreign bonds. The second term on the right is an indirect effect occurring through $\partial M^*/\partial R^*$, which arises because an increase in R^* also increases the opportunity cost of holding foreign money. As these balances decrease, the public has to increase domestic money balances in order to arrive at the desired level of the overall media of payments needed to finance its expenditures.

The direct effect is primarily determined by portfolio selection, which treats domestic money as an asset held for its return relative to other assets. Except in conditions where the domestic bonds do not exist or their security is doubtful, the significant portfolio switch caused by a rise in R^* is likely to be between foreign bonds and domestic bonds (rather than domestic money) since both are income-earning assets. It is not likely to be between foreign bonds and domestic money. Therefore, while portfolio selection analysis implies that $(\partial M^d/\partial R^*)|_{OM^* 0} \leq 0$, this direct effect is likely to be quite weak in normal financial conditions.

The indirect effect in (67) is the multiple of two elements, $\partial M^d/\partial(M^*/p)$ and $(\partial M^*/p)/\partial R^*$. On the second of these two elements, both the portfolio selection and the transactions demand analyses imply that $(\partial M^*/p)/\partial R^* < 0$. On the first element, $\partial M^d/\partial(M^*/p)$ is the substitution between domestic and foreign monies, which constitutes CS. Our earlier discussion on this point implies that $\partial M^{d}/(M^{*}/p) \le 0$. Hence, in (67), the second term on the right is non-negative. Therefore:

$$\Sigma \frac{\partial M^{d}}{\partial R^{*}} \Sigma_{M^{*}=M^{*}} \le 0, \tag{68}$$

$$\frac{\partial M^{d}}{\partial(M^{*}/p)} \le 0 \text{ and } \frac{\partial M^{*}/p}{\partial R^{*}} \le 0 \tag{68^{J}}$$

so that

$$\frac{\partial M^{d}}{\partial(M^{*}/p)} \frac{\partial(M^{*}/p)}{\partial R^{*}} \ge 0 \tag{68^{JJ}}$$

Since, from (68), the first term (the direct effect) on the right-hand side of (67) is non-positive and the second one (the indirect effect), from (68^{JJ}) is non-negative, the sign of $\partial M^{d}/\partial R^{*}$ in (67) – and hence of $\alpha_{R^{*}}$ in (64) – is analytically indeterminate and will depend on the relative magnitudes of the direct and indirect effects. Assuming that the major substitution in portfolio selection is unlikely to occur between the domestic money and foreign bonds,[45] $(\partial M^{d}/\partial R^{*})|_{OM^{*}=0}$ in (67) is likely to be relatively small, so that our hypothesis is that the second term on the right-hand side of (67) will dominate the sign and magnitude of $\partial M^{d}/\partial R^{*}$ for most economies.

Now focusing on the magnitude of $\partial M^{d}/\partial(M^{*}/p)$, our earlier discussion of this CS effect implies that it will be relatively weak in economies in which the foreign money is not extensively used as a medium of payments in domestic transactions.[46] However, in economies in which foreign monies function as one of the domestic media of payments, $\partial M^{d}/\partial(M^{*}/p)$ will be negative and significant. Therefore, in the context of equation (64), if both the domestic and foreign currencies are widely used as media of payments in the domestic economy and the demand for the overall media of payments is determined by domestic GDP, an increase in the return on foreign bonds could decrease the holdings of the foreign currency in the domestic economy, which need to be balanced by an increase in the domestic money balances. That is, the increase in R^{*} would induce a significant increase in M^{d}, so that the second term in (67) is likely to dominate and $\alpha_{R^{*}}$ in (64) should be positive.

Hence, our hypothesis for partially dollarized economies is that $\alpha_{R^{*}}$ should be positive and significant. By comparison, for economies in which foreign money is not in extensive usage as one of the media of payments, the first term on the right-hand side of (67) is likely to dominate, so that $\alpha_{R^{*}}$ in (64) should be insignificant ornegative.

Estimation of $M/(M^*/\rho)$

The third procedure for estimation of the money-demand function focuses on the estimation of M/M^* or of $M/(M^*/\rho)$. We have argued above that there is a special relationship between M and M^* because of their substitution in domestic payments. This could be captured by a weakly separable preference function over M and M^* (as in Bordo and Choudhri, 1982) or a "monetary services production function" (as in Ratti and Jeong, 1994). For such functions, CS can also be assessed by estimating the function for the ratio $M/(M^*/\rho)$ rather than the demand function for domestic money only. In the general, unrestricted, case, this ratio will also be a function of the explanatory variables in (64), so that the corresponding log-linear equation with all variables in logs would be:

$$M^d/(M^{*d}/\rho) = f(\varepsilon^e, R, R^*, Y, \rho) = \beta_0 + \beta_\varepsilon \varepsilon^e + \beta_R R + \beta_{R*} R^* + \beta_Y Y + \beta_\rho \rho \qquad (69)$$

As pointed out earlier, the approximate form of the IRP hypothesis implies that ε^e, R and

R^*

are linearly related for an economy with a high degree of capital mobility, so that ε^e can be left out of the explanatory variables. β_R^* is likely to be positive for economies in which foreign money is significant as a medium of payments in transactions. However, our conjecture is that $\beta_{R*}\ \underline{0}$ for other economies in which the foreign money does not circulate extensively in domestic payments because, among other reasons, retailers do not give the bank rate of exchange and/or charge commissions.

Other studies on CS

In their estimating equation for CS, Ratti and Jeong (1994) claim to combine a "dynamic monetary services" model with portfolio allocation. Their model[47] implies that the optimal ratio $M/(M^*/\rho)$, under purchasing power parity and interest rate parity, equals $(\rho \cdot P/P^*)(R/R^F(1 + \varepsilon))$, so that, in log-linear term:

$$M/(M^*/\rho) = \beta_1(\rho \cdot P/P^*) + \beta_2(R/R^F(1 + \varepsilon)) \qquad (70)$$

where P^* is the foreign price level, $(\rho \quad P/P^*)$ is the real exchange rate, which is included since foreign money balances need to be converted to their purchasing power over domestic commodities, and $(R/R^F(1 \ \varepsilon))$, with $R^* \ R^F(1 \ \varepsilon)$, is the relative rate of return on domestic and foreign bonds. Note that if absolute purchasing power parity (PPP) holds, $(\rho \quad P/P^*) \qquad 1$, so that, in (70), $M/(M^*/\rho) \qquad (R/R^*(1 \qquad \varepsilon))$, which implies that, in a regression, β_1 should not be significantly different from zero. If it is, the PPP hypothesis is rejected. But if IRP holds, then $(R/R^*(1 \qquad \varepsilon))$ equals unity, so that $M/(M^*/\rho) \qquad (\rho \quad P/P^*)$, implying that β_2 should not be significantly different from zero; if it is, the PPP hypothesis is rejected. If both PPP and IRP hold, then $M/(M^*/\rho)\ 1$, implying that both β_1 and β_2 should not be significantly different from zero. Further, if neither PPP nor IRP holds, (70) implies that the coefficients of both $(\rho \quad P/P^*)$ and $(R/R^*(1 \qquad \varepsilon))$ should be unity. Therefore, (70) seems to represent a very restrictive model, whose value lies not so much in providing estimates of CS but rather whether or not either or both PPP and IRP hold.

In economies in which domestic residents have limited or no access to foreign bonds, the return on foreign bonds will not enter the domestic money demand function, so that multicollinearity between changes in the expected exchange rate and returns on domestic and foreign bonds will not pose a problem. This makes the estimation of CS, and the evaluation of its extent, by the estimated coefficient of the expected exchange rate change more credible. For such a context, De Freitas and Veiga (2006) study CS in the context of six Latin American economies.[48] They use a stochastic dynamic optimizing model in which money reduces the losses due to frictions in commodity exchanges. They report evidence of CS for Colombia, the Dominican Republic and Venezuela but not for Brazil and Chile, with ambiguous results for Paraguay.

Conclusions

This chapter has examined the form of the money demand function to be used for estimation. One of these uses expected income as its scale variable, while another uses permanent income. Neither is observable, so that a procedure has to be adopted for their estimation. The rational expectations hypothesis (REH) was proposed for the estimation of expected income, while the adaptive expectations hypothesis was proposed for the estimation of permanent income. Of these, adaptive expectations are backward looking and ignore information that may already be available on the future, but do provide a better measure of permanent income, which is the average expected level of income for the future rather than expected income for the next period.

This chapter also looked at the use of partial adjustment models (PAMs). These models are based on the notion that there are various costs of adjusting money balances quickly, and imply the specific order of the partial adjustment model. The general autoregressive distributed lag model nests the PAM and the adaptive expectations models. An alternative to such a model would be a PAM model with a separate procedure for the rational expectations estimation of expected income.

The open-economy form of the money demand equation distinguishes between currency substitution (i.e. substitution between the domestic and foreign currencies) and capital mobility, which is mainly substitution between domestic and foreign bonds. There are basically three procedures for estimation of currency substitution. These are the estimation of a money demand function, a variable elasticity of substitution function and estimation of the ratio of domestic to foreign currency holdings.

While portfolio theory seems to imply that there should be considerable and increasing currency substitution among the highly open modern economies, the econometric evidence remains quite mixed. This could be due to the preferred habitat role of domestic money balances as the domestic medium of payments. For most economies, foreign monies do not commonly circulate in the economy because of “brokerage costs” imposed by retailers on payments in foreign currencies. However, these costs tend to be trivial for a specific foreign money, which is often the US dollar, in partially dollarized economies, so that such a money functions as a domestic medium of payments in addition to the domestic money. In this case, there should be a high degree of substitution between domestic money and

[24] These authors start with an intertemporal relative risk aversion utility function over consumption in different periods and assume that the time spent shopping per unit of consumption expenditures depends on the domestic and foreign monies held.

foreign money.[49] Note that a fully dollarized economy will not have a distinct domestic money.

9 The demand function for money

Estimation problems, techniques and findings

This chapter presents the estimating function for money demand, an introduction to the appropriate econometric techniques and a summary of the empirical findings on money demand. On the econometric techniques, a major part of the presentation is on cointegration techniques with error-correction modeling for estimating the short-run and the long-run demand for money.

The empirical evidence clearly confirms the dependence of the demand for money on both a scale variable and an interest rate. The issue of which scale variable should be used – current income, permanent income or wealth – is still not settled.

Key concepts introduced in this chapter

- Multicollinearity
- Serial correlation
- Stationarity
- Order of integration
- Unit roots
- Cointegration
- Error-correction modeling

The preceding chapters specify the theoretical analyses of money demand and the general nature of the money demand function. This chapter examines the econometric problems and techniques associated with its estimation, and presents the findings of some of the relevant empirical studies.

A very large number of empirical studies on the money demand function have been published. It would take up too much space to review even the more important of these studies, or to do justice to the ones from which we adopt the results. Among the many excellent reviews of these studies in the literature are those by Cuthbertson (1991), Goldfeld (1973), Feige and Pearce (1977), Judd and Scadding (1982), Goldfeld and Sichel (1990), Miyao (1996) and Sriram (1999, 2000). We shall present only the generic findings on the more significant issues, especially those on the income and interest rate elasticities and the appropriate measure of the monetary aggregate. We intend to pay particular attention to the findings from studies using cointegration and error-correction analysis.

The empirical findings on monetary aggregation reported in Chapter 7 complement the material in this chapter. In particular, the empirical findings concerning the Divisia and certainty equivalence aggregates versus simple-sum aggregates are to be found in Chapter 7 rather than this chapter.

Section 9.1 presents a historical review of money demand estimation and its findings. Sections 9.2 to 9.7 discuss some of the econometric problems that can arise with the data and present the cointegration and error-correction techniques. Section 9.8 presents the findings of some empirical studies using these procedures. Section 9.9 touches on causality. Section 9.10 provides an illustration of the shifts in income and interest-rate elasticities due to financial innovations. Section 9.11 focuses on the desperate search for a stable money demand function.

Historical review of the estimation of money demand

By the end of the 1960s, the basic form of the money demand function had evolved as:

$$m^{\mathrm{d}} = a_0 + a_R R + a_x x \tag{1}$$

where x is a scale variable. The stochastic form of this function was estimated in either a linear or a log-linear form. During the 1960s, the main disputes were whether money should be defined as M1, M2 or by a still wider definition, whether the interest rate should be short-term or long-term, and whether the scale variable x should be income, permanent income or wealth. The data usually used for estimation was annual.

The 1970s were a period of increasing deregulation of the financial system, with financial institutions offering a variety of interest-bearing checking accounts and checkable savings accounts. There was increasing use of quarterly data in this decade and of the partial adjustment model discussed in Chapter 8. The latter justified the use of the lagged value of money among the explanatory variables, so that the linear or log-linear form of the commonly estimated money-demand function was:

$$m_t^{\mathrm{d}} = a_0 + a_r r_t + a_y y_t + (1 - \gamma) m_{t-1} + \mu_t \tag{2}$$

where γ was the adjustment parameter and μ was a white-noise disturbance term.

In an attempt to eliminate serial correlation, a common problem, in money-demand functions or to incorporate a partial adjustment model, (2) was often estimated in its first difference form. The empirical estimates still indicated the stability of the money demand function, but M1 now often, though not always, performed better than M2 and broader aggregates. The value of the adjustment parameter γ in (2) tended to be roughly between 0.20 and 0.5, so that full adjustment to long-run values occurred in about two to six quarters. There was a low impact (one-quarter) real income elasticity (about 0.2) and long-run income elasticity less than 1 (often around 0.7), and a low impact interest elasticity (about −0.02 or smaller) and a long-run interest elasticity roughly between −0.05 and −0.15.[1] The empirical findings on the income and interest elasticities of money demand in Canada were roughly similar.

Income and wealth in the money demand function

The period of the 1950s and early 1960s in many countries was one during which the regulatory authority did not allow interest to be paid on demand deposits. Further, the interest rates paid on savings deposits were subject to upper limits, savings deposits could not be drawn upon by check and a switch from savings deposits to demand deposits often required a personal visit to the relevant financial institution. Under these conditions, the general finding among the empirical studies was that M2 did better than either M1 or measures broader than M2. The explanatory variables that usually performed best with M2 as the dependent variable were medium- or long-term interest rates, with wealth or permanent income as the scale variable. The estimating function was normally stable.

For the data covering the 1950s and 1960s in the USA, regression analysis of the demand for money from equations containing both income and wealth, as well as from equations containing only one of these variables, showed that wealth provided a more stable demand function for money than current income and that, when both variables were included simultaneously, the coefficient of the income variable was insignificant. Permanent income similarly performed better than current income. These results held especially if money was defined as M2 or M3 but not as often in studies where the dependent variable was M1. Further, among functions using income, non-human wealth and permanent income, the empirical estimates showed that functions using a wealth concept gave more accurate predictions of the velocity of circulation of money broadly defined – but not as often for M1– than did those containing current income.

The findings on the economies of scale were uneven. Studies using M1 as the dependent variable often found income elasticities to be less than one, typically around 0.7 or 0.8. Higher income elasticities were usually reported for M2, with some in excess of unity. The reason for this divergence hinges on the inclusion of interest-earning savings deposits in M2. The demand for savings deposits is likely to reflect more strongly a portfolio demand than does M1, so that, with income and wealth positively correlated, the income elasticity of M2 will tend to capture to a greater extent the impact of wealth on savings deposits than does the income elasticity of M1. This portfolio demand could make them a “superior good” for households who experience wealth increases during the sample period.

As between the partial adjustment model and adaptive expectations, with US annual data for 1915–63, Feige (1967) used permanent income as the scale variable and reportedstantaneous adjustment. However, Goldfeld (1973), with quarterly US data, found less than instantaneous adjustment. In general, during the 1970s, studies using quarterly data provided evidence of both adaptive expectations and partial adjustment.

Interest rates in the money-demand function

There are many interest rates in the economy, ranging from the return on savings deposits in banks and near-banks to those on short- and long-term bonds. Near-money assets such as savings deposits in commercial banks proved to be the closest substitutes for M1, so that their rate of return seems to be the most appropriate variable as the interest cost of using M1.

But if a broader definition of money were used, the interest rate on medium-term or long-term bonds would become more appropriate (the alternative to holding M2 or M3 is longer term bonds), since the savings components of the broad definition of money themselves earn an interest rate close to the short rate of interest.

The interest rates usually used in estimating money demand are: the interest rate paid on savings deposits in commercial banks, or on those in credit unions (such as Mutual Savings Banks and Savings and Loan Associations in the USA, Caisses Populaires in Quebec, Canada); the yield on Treasury bills or on short-term prime commercial paper and the yield on longer term bonds, such as 3 to 20-year government or commercial bonds. Each of these interest rates seems to perform fairly well, sometimes better and sometimes worse than others, in some study or other, and yields different coefficients.

A uniformly good performance, irrespective of which of the interest rates is included in the regression, is an indication that the various interest rates are related, moving up or down in a consistent pattern, so that it is immaterial which interest rate is included. One theory that points towards such consistency of pattern is the expectations hypothesis on the term structure of interest rates, i.e. on the yields on assets differing in maturity. Chapter 20 presents this hypothesis for the financially well-developed financial markets, pointing out that it has done remarkably well in explaining the differences in the yields of assets differing in maturity. A consequence of such a relationship among interest rates is that the inclusion of more than one interest rate results in multicollinearity and therefore in biased estimates of their coefficients.

However, while the relevant interest rates are closely related, they do not move so closely together that any of them will do equally well in estimation, so that usually one or two of them have to be chosen on empirical grounds for inclusion as regressors. On the wider question of whether the demand for money depends on interest rates or not, there is substantial evidence that the demand for money does depend negatively upon the rate of interest in financially developed economies. This is also the finding of many studies on the less developed countries (LDCs).

Some studies on the LDCs, however, do not find significant interest rate elasticities for a variety of reasons, including regulatory limits on the interest rates in the economy and inadequate access to banking and other financial facilities. In these cases, very often the rate of inflation rather than the published data on interest rates yields better empirical results. This occurs because the regulated interest rates usually do not accurately reflect the expected rate of inflation, as market-determined rates do in developed financial markets, so that land, inventories and other real assets, whose prices better reflect the rate of inflation, become more attractive alternatives than bonds for holding cash.

Various empirical studies have reported that the interest elasticity of money demand is definitely negative and significant, and in the range -0.15 to -0.5.

Money demand and the expected rate of inflation

One of the alternatives to holding money is commodities, which have the (expected) rate of return equal to the expected rate of inflation less their storage and depreciation costs. Some of the commodities – as for example, untaxed plots of land – have minimal storage and depreciation costs, so that the (expected) rate of return on commodities is usually taken to be proxied by the expected rate of inflation. Therefore, the expected rate of inflation is one of the arguments in the money-demand function, in addition to interest rates, as Friedman's analysis of money demand in Chapter 2 pointed out.

However, in perfect financial markets, for small values of the real interest rate and expected inflation, the nominal and the real rates of interest are related by the Fisher equation:

$$R_t = r_t + \pi^e_t \tag{3}$$

where R is the nominal rate of interest, r is the real one and π^e is the expected inflation rate. At significant rates of inflation, variations in the real rate tend to be much smaller in magnitude than the expected inflation rate, so that R_t and π^e_t will be closely correlated. Given this close correlation and that between π^e_t and the actual rate of inflation π_t , R and π also tend to be closely correlated in periods with significant inflation rates. Therefore, incorporating both R_t and π_t in the money demand equations often leads to multicollinearity and biased estimates of their coefficients. As a way around these statistical problems, π_t is often dropped in favor of R_t from the estimated money demand equations for developed economies with market determination of R_t. However, economic theory implies its inclusion in addition to the inclusion of interest rates, so that its omission could result in a misspecified equation.

In economies such as the LDCs', where the financial markets are not well developed, ceilings are often imposed on the rates of interest that can be legally paid and there could exist both an official interest rate and a free or black market rate. Further, reliable data on interest rates may not be available. In these cases, π^e_t should be retained in the estimating equation in addition to – and sometimes even to the exclusion of – the interest rate. Note that the proper variable is π^e_t and that π_t is only one of the possible proxies to it.[2]

The liquidity trap

One of the questions of interest in monetary theory since the time of Keynes, discussed in Chapter 2 above, has been about the empirical existence of the liquidity trap. Keynes posited the possible existence of such a trap, though he also expressed the belief that he did not know of any case where it had existed.

One possible method of testing for the existence of the liquidity trap is to estimate the demand for money separately for periods with differing ranges of the prevailing interest rates. Estimates showing that the interest elasticity of demand tends to increase in periods with lower ranges of interest rates, and especially those showing a substantial increase at very low interest rates, can be interpreted as raising a presumption that the liquidity trap could have existed empirically. However, empirical studies so far have not revealed such a pattern. Velocity functions estimated separately for each decade did not find any higher interest elasticity of the demand for money during the 1930s, when interest rates were low than during other decades with higher interest rates. Further, regressions incorporating data from the 1930s did fairly well in predicting velocity during the subsequent decades, implying that the interest elasticities during the 1930s did not differ substantially from those of more normal conditions. These studies indicate that the liquidity trap does not seem to have existed in the US economy for any significant period, if at all, during the Great Depression of the 1930s and is even less likely a possibility for other periods.

Theoretically, the liquidity trap should come into existence if the nominal yield on bonds becomes zero. The Japanese economy in recent decades provides an interesting experiment on the liquidity trap since it has had short-term interest rates close to zero. Bae *et al.* (2006) have studied different money demand functions for Japan using linear and non-linear cointegration techniques with quarterly data from 1976:1 to 2003:4. They report that the interest elasticity for their various monetary aggregates, including M1, is much higher at low interest rates than at higher rates, thereby favoring the conjecture that the liquidity trap may exist at interest rates that are zero or close to zero.

Shifts in the money demand function

Much greater impact of financial deregulation was felt in the 1980s than had been permitted or achieved in the 1970s. Further, technological and product innovation in the financial sector was very rapid. Computers also came into general use in firms and households, and permitted more efficient management of funds. By the end of the 1980s, automatic tellers for electronic transfer and withdrawal of funds from both demand and savings accounts had become common, and were more numerous than bank branches. Many new variants of demand and savings deposits had been created and the distinction between demand and savings deposits in terms of their liquidity became blurred almost to the point of disappearance, though savings deposits still paid higher interest but also imposed higher charges. Deregulation, innovation and technological change resulted in a failure of the quarterly specification for money demand, whether money was defined narrowly or broadly.

These developments in the 1980s led to the estimated demand functions performing poorly, with unstable money demand and with a highly variable velocity of circulation. The econometric tests also became much more sophisticated than in earlier periods. Among these, econometric tests of the money and income time series showed that they were not stationary. To deal with this, cointegration analysis became one of the preferred techniques and showed that money and income were indeed cointegrated, as were interest rates with them over many periods.

Common problems in estimation: an introduction

This section is intended to show that the estimation of the money-demand function is not a simple and straightforward matter, and that application of the classical least-squares regression technique to its estimation need not provide reliable estimates. The section provides a brief treatment of the common problems encountered in money demand estimation and is not meant to provide a complete, in-depth or rigorous treatment of the econometric problems discussed or of the appropriate econometric techniques. These are left to econometrics textbooks such as Davidson and Mackinnon (1993).

The general form of the demand function for real balances implied by the transactions, speculative, buffer stock and precautionary analyses is of the type:

$$M^{d}/P = m^{d} = m(R_1,\ldots, R_m,\pi^{e}, y, w) \tag{4}$$

where:

M = nominal balances
m = real-money balances
P = price level
π^{e} = expected rate of inflation
R_i = rate of return on ith near-money asset, $i = 1,\ldots, m$
y = real income/expenditures
w = real wealth.

The following subsections consider some of the econometric issues that arise in the estimation of such a money demand function.

Single equation versus simultaneous equations estimation

From a *general equilibrium* viewpoint, the rate of return on each of the near-money assets is influenced by the demand and supply of money and also by the demands and supplies of risky assets. A general empirical study of the demand for money would then simultaneously estimate the demand and supply functions for all the financial assets, where the demand function for the *i*th asset is:

$$x_i = x_i(R_1,\ldots,\ R_m, R_{m+1},\ldots,\ R_n, \pi^e, y, w) \qquad (5)$$

The definitions of the symbols are:

$x_i =$ real quantity of the *i*th monetary asset, $i = 1,\ldots, m$

$R_j =$ rate of return on the *j*th non-monetary asset, $j = m + 1,\ldots, n.$

Note that from a rigorous general equilibrium viewpoint, each asset should be homogeneous. A general equilibrium study becomes an extremely large enterprise and poses its own econometric problems. Most studies of the demand for money have been partial and, for statistical and other reasons, have used various degrees of aggregation in defining money. They also often confine the explanatory variables to one rate of interest and either income or expenditures or wealth. However, whether or not one is estimating the demand functions for several assets simultaneously, it is important to consider the cross-equation restrictions that the relevant theory might imply for them. We illustrate these in the following for the case of the allocation of a portfolio between money and bonds, as analyzed in Chapter 5.

Estimation restrictions on the portfolio demandfunctions for money and bonds

Chapter 5 implied that the general form of the speculative demand functions for assets is:

$$x_i^d = x_i^d(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, W) \qquad i = 1,\ldots,\ n-1 \qquad (6)$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\rho}$ are respectively the vectors of the mean returns, the standard deviations and the correlation coefficients among the values of the assets, and *W* is the wealth allocated among the assets. (6) and the portfolio budget constraint imply that:

$$x_n = W - \Sigma_i x_i^d(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, W) \qquad i = 1,\ldots,\ n-1 \qquad (7)$$

so that the demand function for one of the assets must be derived as a residual from the estimated demand function of the other assets. Alternatively, if the demand functions for all the assets are being estimated, the appropriate cross-equation restriction must be imposed on the estimating equations. As an illustration of this, the restrictions imposed by (7) for the two-asset case of money (*M*) and the composite bond (*B*) are set out below.

Suppose the estimating equations for *M* and *B* are linear and are specified as:

$$M = a_0 + a_1 R_m + a_2 R_b + a_3 W \qquad (8)$$

$$B = b_0 + b_1 R_m + b_2 R_b + b_3 W \tag{9}$$

where R_m is the nominal return on money and R_b is the nominal return on bonds. The budget constraint on M and B is:

$$M + B = W \tag{10}$$

Substituting (8) and (9) into (10) yields:

$$(a_0 + b_0) + (a_1 + b_1)R_m + (a_2 + b_2)R_b + (a_3 + b_3 - 1)W = 0 \tag{11}$$

To satisfy (11) for all possible values of the variables, each term in it has to be zero. Therefore, we must have:

$$a_3 + b_3 = 1 \tag{12}$$

$$a_i + b_i = 0 \qquad i = 0, 1, 2 \tag{13}$$

Failure to impose these restrictions on the estimated coefficients in the simultaneous estimation of both demand functions will generally yield estimated values of the coefficients that are not consistent with the budget constraint and are therefore not valid. In cases where a single demand function, say for money, is estimated, and its estimated coefficients seem to be quite plausible, the *implied* values of the coefficients for the bond equation may not prove to be accurate or even plausible. For example, if the estimated elasticity of the demand for money is much larger than one, this would in turn imply that the elasticity of the demand for all other financial assets is correspondingly less than one, which may not be plausible for the economy and the period in question, thereby leading to a rejection of the estimated money demand function. Therefore, if it is feasible, it would be better to estimate *simultaneously* the complete system of demand equations and impose appropriate restrictions on the coefficients. However, this is not always feasible and often exceeds the researcher's interests, so that most studies tend to confine themselves to the estimation of only the money-demand function.

The potential volatility of the money demand function

Note that the coefficients a_i, $i = 0$, 1, 2, 3 in the money-demand function (8) depend upon the means, the standard deviations and the correlation coefficients of the expected terminal values of the assets, for all of which the subjectively – not objectively – expected future (not the past actual) values are the relevant ones. If these characteristics of assets change, the implied coefficients will change and the demand functions will shift. In the real world, subjective expectations on the returns and future values of the financial assets continuously shift for a variety of reasons, so that the subjectively based characteristics of assets are constantly changing. These sources of shifts can be classified into (i) shifts in subjective probability estimates because of changing market conditions, (ii) shifts in policies which shift the outcomes and their probability functions, and (iii) innovations in the payments mechanism, such as the introduction of ATMs and electronic banking.

Keynes (1936, Ch. 13) focused on (i) and argued that the expectations of asset returns, and hence of these characteristics, are very volatile. This argument implies that the demand functions for money and other financial assets would be constantly shifting, so that they could not be properly estimated or, if estimated, would be worthless – unless the nature of the shift could be specified and adjustments made for it – as guides for future policies.

The Lucas critique (in Chapter 17) of estimated functions used for policy purposes focuses on (ii) above and argues that, if a change in policy – for example, in the monetary regime, tax laws, banking and financial regulations, relevant political stance, etc. – shifted the characteristics of the returns to the assets, the demand functions would shift and the prior estimated forms would no longer be valid. Hence, specific forms of the demand functions will not hold across policy regimes.

The above arguments caution that, since the money demand and supply functions, as well as other relevant policy functions, are constantly changing and definitely do so over decades, the validity of using data over long periods of time to estimate a demand function with constant coefficients should be extremely suspect. This is especially so in a period of financial innovation, which keeps changing the relative characteristics of the existing assets and, over time, keeps adding newer ones to the marketplace.

Multicollinearity

Another statistical problem encountered in partial studies is the *multicollinearity problem*. Suppose that the demand for money is related to both income and wealth but that income and wealth are themselves highly correlated. The estimate of the relationship between money balances demanded and income is then influenced by the relationship between income and wealth and vice versa, so that the estimated relationship may not be an accurate measure of its actual value.

Similarly, the various rates of return are highly correlated, so that the estimates of the coefficients of the rates of return in the money demand function in the economy also tend to be biased and must be treated with caution.

If there is fairly close correlation among a set of variables, one solution to the multicollinearity problem is to use only one of the variables in the set and interpret its estimated coefficient as representing the collective effect of all the variables in the set. For instance, given the close correlation among the interest rates, most money demand functions include among the independent variables only one interest rate in order to avoid multicollinearity. This is usually a short-term rate, such as the Treasury bill rate. However, some studies include both a short-term and a long-term interest rate. As between current income and permanent income or wealth, while some studies include only current income, others include permanent income, with multicollinearity between these two variables preventing the simultaneous inclusion of both.

Serial correlation and cointegration

Most regression techniques assume that the error terms are serially uncorrelated and have a constant variance. These should be checked for the estimated error. If it does not satisfy these conditions, as often proves not to be so, the estimated coefficients will be biased and the appropriate techniques that can ensure unbiased estimates have to be used. The techniques often used for correcting for serial correlation include estimating the money demand function in a first-difference form and using a technique with a built-in correction for the relevant order of serial correlation.

Regression analysis used for deriving the money-demand function assumes that the variables are *stationary*. A variable is not stationary if it has a trend or/and serial correlation. Many of the variables in the money demand function, such as income and the money stock, are not stationary. If this happens, the use of classical regression techniques, such as one-stage least squares, two-stage least squares, etc., yield biased estimates of the coefficients of the independent variables. The preferred procedure in such cases is that of *cointegration analysis*.

The relationship between economic theory and cointegration analysis: a primer

This section presents a brief introduction to stationarity and cointegration analyses. The reader is referred to econometric textbooks for a proper treatment of these topics.

Economic theory: equilibrium and the adjustment to equilibrium

An economic theory is intended to explain the determination of the actual values of a selected economic variable or of several economic variables. As the starting point for the following exposition, we focus on the determination of a single economic variable, say *y*. The theory on its determination examines three questions:

1 Does an equilibrium relationship exist between the dependent variable *y* and its explanatory variables $\boldsymbol{x}$ $(=x_1,\ x_2,..\ .)$? Suppose that such a relationship exists and is of the form:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \qquad (14)$$

where x_0 is taken to be constant, and *y* and $\boldsymbol{x}$ could be the levels of the variables, their first differences or rates of change, etc., or some mix of these, and the relationship could be linear or non-linear. For the following exposition, the equilibrium relationship is assumed to be among the levels (or the log values) of the variables and to be a linear one. The estimation equation, expanded to include lagged values of the dependent variable *y* and the explanatory variables $\boldsymbol{x}$ on the right-hand side, becomes an autoregressive distributed lag (ARDL) equation, whose treatment is presented in the Appendix to this chapter.

If all the variables in the relationship are stationary, classical least-squares techniques can yield unbiased estimates of the coefficients of the variables. However, if some of the variables in the relationship are not stationary, these techniques do not yield unbiased estimates. The cointegration estimation technique is likely to yield better results. To determine the technique that is appropriate, the stationarity or otherwise of each of the variables has first to be determined by stationarity tests. These are discussed later.

2 Is the equilibrium unique? It is assumed that there is a unique equilibrium for the given structural specification of the equations of the model.

3 Is the equilibrium relationship between *y* and $\boldsymbol{x}$ stable or unstable? Assuming it to be stable, there would be a dynamic adjustment path during the disequilibrium following a disturbance to the equilibrium relationship. The dynamic path can be of different types, requiring different specifications of the adjustment process.[3] It is often not clear which one is the empirically relevant process for the structure in question. The most common assumption is that the adjustment process is linear (or log-linear). The estimation of the

adjustment process and its implications for the stability of equilibrium are discussed later under the heading of error-correction models (ECMs).

Stationarity of variables: an introduction

The equilibrium relationship between the endogenous variable *y* and the vector ***x*** of explanatory variables was specified above as:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{15}$$

Estimation of this equilibrium relationship by classical regression techniques requires that each of the variables be stationary.

A variable z_i is said to be stationary if its mean, variance and covariances with the other variables in the relationship are finite and constant. The usual symbols for these are:

$$E(z_i) = \mu_i$$

$$V(z_i) = \sigma_{ii} = \sigma_i^2$$

$$\mathrm{COV}(z_i, z_j) = \sigma_{ij}$$

The stationarity of z_i implies that these moments of its distribution will remain unchanged, except for random differences, over different sample periods. Conversely, if estimation over different sample periods yields different estimated values of these moments, the variable is likely to be non-stationary. If any of the variables in the relationship implied by the theory is non-stationary, then the estimates obtained by classical least-squares regressions of the equilibrium relationship among the variables will differ among the various sample periods, so that the estimated relationships will not accurately reveal the true relationship.

Causes of non-stationarity

The potential causes of non-stationarity are:

1 The mean value of the variable is not stationary, due to a trend.

2 The variance of the variable and its covariances with other variables are not stationary. This is due to serial correlation.

If the variables in the estimation are not stationary due to serial correlation, two different types of estimation procedures can be attempted for estimating the true equilibrium relationship. One of these is to render the data series stationary prior to estimation, such as by employing a procedure for eliminating serial correlation. To render a series (with serial correlation) stationary, each would be differenced once or more times until its derived series is stationary. Alternatively, a correction for serial correlation, such as the Cochrane–Orcutt method, can be used in the estimation process. Classical regression techniques, such as one-stage or two-stage least squares, often employ such procedures to deal with non-stationary time series.

An alternative to the above procedure is to use the following property of the equilibrium relationship, with y as the dependent variable and, for illustration, only x_1 and x_2 as the explanatory ones. Assume, as before, that the equilibrium relationship is linear in the levels (or log values) of the variables, so that it is of the form:

$$y_t = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \tag{16}$$

which can be rewritten as:

$$y_t - \alpha_0 - \alpha_1 x_1 - \alpha_2 x_2 = 0 \tag{17}$$

In this equation, since the right-hand side (which is zero) is stationary, the composite variable ($y_t \quad \alpha_0 \quad \alpha_1 x_1 \quad \alpha_2 x_2$) on the left-hand side must also be stationary. Hence, while the individual variables are not stationary, their linear combination given by (17) would be stationary. The appropriate linear combination is one with the coefficients (1 α_0 α_1 α_2). Note that α_0 is the (coefficient of the) constant term. The vector (1 α_0 α_1 α_2) is called the cointegrating vector (in this case, with the coefficient of the dependent variable normalized to unity) and (17) is called the *cointegrating equation*. Empirical analysis requires an appropriate estimation procedure that will provide unbiased estimates of this vector. A few points about the above relationship need to be noted.

- Multiplying each of the coefficients by a constant yields a stationary variable, so that any multiple of the cointegrating vector is also a cointegrating vector.
- It is quite appropriate to set the coefficient of the endogenous variable as unity, so that it is customary to normalize the cointegrating vector in this way.
- The signs of the coefficients of the explanatory variables in the cointegrating vector and equation are the reverse of those in the equilibrium relationship.
- If the equilibrium relationship is linear in the logs of the variables, then the cointegrating vector will specify (with signs reversed) the elasticities of the endogenous variable with respect to the explanatory variables.

The next section discusses the sources of non-stationarity of the variables and the estimation procedures for determining whether a variable is stationary or not.

A non-stationary mean due to a trend

A trend in a variable will make its mean non-stationary. An example of this occurs if:

$$z_t = \alpha_0 + \alpha_1 t + \mu_t \tag{18}$$

where z is the variable in question, t is time and μ is white noise.[4] In this case, data samples over different periods will yield different mean values of z, since:

$$Ez_t = \alpha_0 + \alpha_1 Et$$

A variable that is nonstationary because of the presence of a trend can be transformed by removing the trend into a corresponding variable (i.e. $z_t \; \alpha_1 t$) that is stationary. Such a variable is said to be trend-stationary (TS).

Non-stationary variances and covariances because of serial correlation

This type of non-stationarity arises if the variable behaves according to:

$$z_t = \alpha_0 + z_{t-1} + \mu_t \qquad (19)^{5}$$

where μ_t is white noise. z_t is said to follow a random walk if the constant term α_0 is zero; it follows a random walk with drift if α_0 is not zero. The value of z_t depends on the actual value of $z_{t\ 1}$ (which includes the actual value of $\mu_{t\ 1}$), so that there is a stochastic tendency for the mean of the variable to change over different data samples.

Rewrite this equation as:

$$Oz_t = z_t - z_{t-1} = \alpha_0 + \mu_t \qquad (20)$$

where Oz_t (z_t $z_{t\ 1}$) is stationary. Therefore, a variable that is non-stationary because it follows a random walk can be rendered stationary by taking its first difference. Such a variable is called difference-stationary (DS). If taking the first difference of a series makes it stationary, it is said to be integrated of order 1, which is written as I(1).

Note that if a series is I(1), taking its first difference will yield a stationary series. But if the series also has a time trend, the first difference of the series will still possess a trend, so that it will not be trend-stationary. An adjustment for this trend will have to be made in the estimation procedure.

Non-stationarity because of a shift in the value of the variable

Note that a variable may be stationary but that its data sample may indicate non-stationarity because of a shift at some point in its time series. In this case, classical least-squares can still be used with the shift captured through the use of a dummy variable.

Order of integration

In the general case of serial correlation, a variable may follow the process:

$$z_t = \alpha_0 + z_{t-p} + \mu_t \qquad (21)$$

In this case, the variable would have to be differenced p times to arrive at a stationary series. The variable is then said to be integrated of order p and is designated as I(p).

In the case of difference-stationary data, while using the appropriate number of differences of the variables does eliminate the problems posed by the non-stationarity of the levels of the variable, a regression using differenced data eliminates the relationship among the levels of the variables, so that the regression will not provide estimates of the long-run relationship between the *levels* of the dependent and the independent variables in the estimating equation. Therefore, the use of differenced data is not a proper strategy for finding the equilibrium relationship among the levels of the variables. For example, in the context of the money-demand function, the underlying theory implies an equilibrium relationship between the levels of the variables, so that using differenced data will not provide an estimate of this function.

Note that if a series is I(p), taking its pth difference will yield a stationary series. But if the series also has a time trend, the pth difference of the series will still possess a trend, so that it will not be trend-stationary and an adjustment for this trend will have to be made in the estimation procedure.

Testing for non-stationarity

Since non-stationarity can arise from both a trend and serial correlation, the appropriate test for stationarity has to simultaneously test for both these. The following discusses such tests.

Suppose that the variable z follows an autoregressive, non-stationary, data-generating process with a one-period lag:

$$z_t = a_0 + a_1 t + a_2 z_{t-1} + \mu_t \tag{22}$$

where t is time and μ_t follows a stationary process. Subtracting z_{t-1} from both sides,

$$Oz_t = a_0 + a_1 t + (a_2 - 1)z_{t-1} + \mu_t \tag{23}$$

If a_2 1, z_t is I(1). The test for a_2 1 as against $a_2 < 1$ is called a unit root test. Such a test is referred to as the Dickey–Fuller (DF) unit root test. The estimation of this equation can yield the following results:

1. If $\hat{a}_0 = \hat{a}_1 = 0$ and $\hat{a}_2 = 1$, then z follows a random walk and its series is I(1).
2. If $\hat{a}_0 \neq 0$, $\hat{a}_1 = 0$ and $\hat{a}_2 = 1$, then z follows a random walk with drift and its series is still I(1).
3. If $\hat{a}_2 = 1$ and $\hat{a}_1 \neq 0$, then z has a trend and is trend-stationary.

A more sophisticated test for the sources of non-stationarity is provided by the Augmented Dickey–Fuller (ADF) unit root test.[6] This test allows for higher-order autoregressive processes and is based on the estimation of the equation:

$$Oz_t = a_0 + a_1 t + (a_2 - 1)z_{t-1} + \sum_{j=1}^{n} b_{ij} Oz_{t-j} + \mu_t \tag{24}$$

which allows for the impact of n lagged values of the variable. The ADF unit root test is for the null hypothesis that a_2 1, against the alternative that $a_2 < 1$.[7] Failure to reject the null hypothesis implies non-stationarity of the series.

If the ADF and other tests[8] for the data series of the variables in a relationship show that at least some of the series are I(p), $p \geq 1$, the relationship has non-stationary variables so that, as mentioned above, the classical regression techniques – such as ordinary least squares – will not provide unbiased and consistent estimates of the coefficients of the relationship. An appropriate technique would be cointegration.

Cointegration and error correction: an introduction

The cointegration technique is based on the assumption of an equilibrium (linear or log-linear) relationship among the variables, which implies that two or more variables that are individually non-stationary but are integrated of the same order possess a linear combination of a one-degree lower order of integration.[9] Therefore, if all the variables are I(1) and are cointegrated, then their cointegrating equation would yield a composite variable of order I(0), i.e. it would be stationary. As explained earlier in the discussion on the connection between an equilibrium relationship and cointegration, if the equilibrium relation among a set of I(1) variables is linear (log-linear), the existence of such a linear (log-linear) combination is the equilibrium relationship implied by the relevant theory. Cointegration techniques attempt to estimate whether such a combination exists and, if so, what is the cointegration vector.[10] The cointegration equation based on such a vector is then treated as an estimate of the long-run equilibrium relationship.

If the variables are all I(p), then their cointegrating vector, if it exists, will yield a variable which is I($p - 1$).

In practice, problems in using cointegration analysis arise if the variables in the relationship implied by the theory are of different orders of integration. If y is I(2) and some of the x_i, $i = 1, 2, \ldots, n$, are I(1) while others are I(2), the successful[11] application of the cointegration technique to the I(2) variables only would yield a cointegrating equation that provides an I(1) composite variable. The I(1) estimate of this composite variable can then be used along with the I(1) variables in the error-correction estimation, discussed later. A similar procedure would have to be used if y is I(1) and some of the x_i variables in the relationship implied by the theory are I(0) while others are I(1).

If the dependent variable is of a lower order of integration than some or all of the explanatory variables implied by the theory, then it is inappropriate to use cointegration analysis. This would occur if y is I(0) (i.e. stationary) while some or all the explanatory variables are I(p), $p \geq 1$.

Estimation problems therefore arise if the variables are of different orders of integration. In such a case, it might be more appropriate to use a cointegration procedure that allows such variability. Pesaran *et al.* (2001) provide such a procedure.

Relationship between cointegration results and economic theory

Let the relationship derived from economic theory be of the form:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{25}$$

If the variables $y, x_1,\dots, x_n$ are all I(1), the general form of the cointegrating vector, if one is found, is:

$$f(y, x_1,\dots, x_n) = 0$$

This relationship is log-linear or linear depending upon whether the data was in logs or not. The form of the *cointegrating equation* is:

$$y_t - \alpha_0 x_0 - \alpha_1 x_1 - \alpha_2 x_2 - \cdot\cdot\cdot - \alpha_n x_n = 0 \tag{26}$$

where x_0 represents the constant term. As mentioned earlier, its signs of the coefficients of the explanatory variables have to be reversed to arrive at the original equation (25).

Because of potential econometric problems, various econometric checks (discussed later) are applied to check the econometric acceptability of the estimated coefficients. However, even if the estimate is acceptable on the basis of the econometric tests, the estimated cointegrating vector may still not be a plausible estimate of the true equilibrium economic relationship. From the perspective of economic theory, this plausibility is judged by checking whether the signs of the cointegrating vector are consistent with those implied by the theory and whether the estimated magnitudes of the normalized cointegrating equation are plausible in terms of the theory, intuition and estimates obtained by other studies. If this is not so, the estimated cointegrating vector will have to be rejected as an estimate of the true equilibrium relationship.

Deviations from the equilibrium relationship: the error-correction assumption for adjustments in disequilibrium

It was assumed earlier that the equilibrium relationship between the endogenous variable and the vector of explanatory variables is stable and unique. Cointegration literature labels this equilibrium relationship – and its estimate by the cointegrating vector – the *long-run* relationship.

Since the equilibrium has been assumed to be stable, any deviations from it will be corrected through an adjustment process.[12] In cointegration analysis, this adjustment process is often referred to as the *dynamic adjustment* or as the *short-run relationship* between the endogenous variable and the explanatory variables. Cointegration techniques assume that the dynamic adjustment follows a linear or log-linear error-correction process, rather than some other one. This process is in the nature of a partial linear adjustment each period.[13]

The cointegration literature refers to its adjustment estimation technique as "the *error-correction model*" (ECM). This model specifies the change in the endogenous variable *y* as a function of last period's error between the actual and the equilibrium value of the dependent variable and of the change in each of the explanatory variables of the equilibrium relationship. Other variables, provided that they are stationary, can also be introduced in the ECM. The linear specification of the error-correction element of the ECM is specified as:

$$Oy_t = \theta(y_{t-1} - y_t{}^*{}_{-1}) + \cdots \tag{27}$$

where Oy_t y_t y_{t-1}, y^* is the equilibrium value (calculated from the estimated cointegrating vector), and y changes each period by the fraction θ of the previous period's deviation of the actual value from the equilibrium value y^*. For the equilibrium to be stable (equilibrium-reverting), we need $\theta \leq 1$. The complete ECM equation would have the form:

$$Oy_t = a_0 - \sum_{i=2}^{p} a_i Oy_{t-i} + \sum_{j=1}^{n} \sum_{i=1}^{q} b_{j,t-i} Ox_{j,t-i} - \theta \mathrm{ECM}_{t-1} + \eta_t \tag{28}$$

where the lag lengths p and q have been optimally determined and:

$$\mathrm{ECM}_{t-1} = y_{t-1} - \hat{\alpha}_0 - \sum_{j=1}^{n} \hat{\alpha}_j \boldsymbol{x}_{jt-1} \tag{29}$$

Cointegration techniques

The two popular cointegration procedures for determining the equilibrium relationship or relationships among non-stationary variables are the Engle–Granger (Engle and Granger, 1987) and the Johansen – also called the Juselius-Johansen – procedures (Johansen and Juselius, 1990; Johansen, 1988, 1991). The most common application of these procedures is when all the variables are I(1).

Engle–Granger method for a reduced-form equation[14]

For the estimation of the cointegration vector and its associated error-correction dynamic adjustment equation, the Engle–Granger method uses a two-stage procedure. In the first stage, it estimates the cointegrating vector among the I(1) variables for a given equilibrium relationship and tests the residuals for stationarity. If these residuals are stationary, as they should be if all the variables are I(1), the second stage uses them to estimate the dynamic short-run response of the dependent variable by the error-correction model.

The Engle–Granger technique is quite appropriate if all the explanatory variables are exogenous. Often, a model has several endogenous variables, so that it possesses several equilibrium relationships among its variables. In this case, the Johansen procedure would be preferable to the Engle–Granger one.

Johansen cointegration procedure for a model with several endogenous variables[15]

In a model where more than one variable is endogenous, there would be more than one equilibrium relationship among the variables. The Johansen cointegration procedure is then the preferable one since it treats all the variables in the estimation process as endogenous and tries to simultaneously determine the equilibrium relationships among them. In addition, this procedure provides estimates of the cointegrating vectors and the error-correction model in one step. These advantages have made the Johansen procedure the more common one in the cointegration literature.

Assuming that all the variables being considered are I(1), the Johansen procedure (a) takes all the I(1) variables to be as if endogenous and related by a vector-autoregressive (VAR) structural model, (b) uses the maximum likelihood estimation for the VAR model, and (c) derives a set of cointegrating vectors. The number of cointegrating vectors is determined by the eigenvalue and trace tests. The maximum number of independent equilibrium relationships that can exist among a set of endogenous variables has to be one less than the number of variables.[16] Therefore, the maximum number of significant cointegrating vectors should be one less than the number of variables in the VAR model.

The propensity of the Johansen procedure to yield several (significant) cointegrating vectors among the variables is an asset but also raises two troublesome issues:

1 Which vector should be treated as the estimate of which one of the equilibrium relationships among the variables? That is, a choice has to be made among the available cointegration vectors for the particular economic relationship being sought. This choice is usually made on the basis of the signs implied by the theory for the coefficients and the estimated magnitudes of the coefficients falling within a plausible range.

2 Any linear combination of the estimated cointegrating vectors is also an admissible cointegrating vector. Therefore, one can generate an infinite number of combinations, many of which are usually likely to fit the requirements of the appropriate signs and magnitudes being sought for a specific relationship. The linear combinations can be searched for this purpose. However, this search can easily degenerate into "vector- mining."

To illustrate, in some applications of the Johansen technique to money-demand estimation, it is found that the elements of none of the cointegration vectors possess signs consistent with the a priori expectations on the elasticities of the money demand function. Alternatively, these elements could be such as to imply implausible magnitudes of the elasticities. These problems could arise from the limited sample size, inaccuracies in the data, misspecification in the set of variables, breaks in the data, etc. But it is also possible to argue that, since a linear combination of the cointegrating vectors is also a cointegrating vector, one could try to find that linear combination of the cointegrating vectors such that the elements have the desired signs and magnitudes in a plausible range. However, this amounts to "mining the vectors," so that the results often fail to convince other researchers.

To conclude, while the Johansen technique provides econometric evidence on the existence of long-run relationships among a set of variables, the identification or derivation of the structural coefficients of the model from the elements of the cointegration vectors can be quite problematical.

Cointegration, ECM and macroeconomic theory

Economic theory often implies more than one long-run relationship among any given set of economic variables. For example, in the IS–LM model, money demand depends upon national income and interest rates, while national income – as do interest rates – depends on the money supply, which equals money demand in equilibrium. Assuming these three variables to be all I(1), such a simultaneous determination of economic variables implies the possible existence of a maximum of two cointegrating vectors among them. In general, for *n* variables, there could be (*n*–1) independent cointegrating vectors. This poses a problem since the cointegration technique does not identify a given cointegrating vector with a specific economic relationship. For instance, suppose two cointegrating vectors are found among money, income and interest rates. The econometric estimation by itself does not make it clear which one of the cointegrating vectors specifies the money demand relationship. This has to be decided by the researcher on the basis of the signs imposed by economic theory on the coefficients of the money demand relationship and on the basis of the plausibility of the magnitudes of the elements in the cointegrating vectors. The elements of the selected cointegrating vector are then taken to specify the respective long-run coefficients of the linear (or log-linear) money demand function.

Now, assuming that a cointegrating vector exists, the ECM can be used to capture the adjustment of the dependent variable to the long-run equilibrium specified by the cointegrating vector. Among the characteristics of the ECM are:

1 It defines the deviation from the long-run value as the “error” and measures it by the residual, i.e. the difference between the actual value of the dependent variable and its estimated value based on the selected cointegrating vector.

2 It specifies the first difference of the dependent variable as a function of this error lagged one period, the I(0) variables and the first differences of the independent I(1) variables.[17] Appropriate lags in the latter are introduced at this stage.

3 The coefficient of the lagged residual is the error-correction coefficient and specifies the speed of adjustment of the dependent variable to its long-run value.

4 The estimated coefficients measure the short-run movements in the dependent variable in response to fluctuations in the independent variables.

Application of the cointegration–ECM technique to money demand estimation

To illustrate the application of the cointegration–ECM procedure to money demand, let the long-run money demand function be:

$$m^{\mathrm{d}}_t = a_0 + a_R R_t + a_y y_t \tag{30}$$

[17] This model is valid only if the estimated error is stationary.

Assume that the data series for *m*, *R* and *y* are all I(1). Let their estimated cointegrating *vector* be $(1\ -a_0\ -a_R\ -a_y)$ in which the second, third and fourth elements have the opposite sign to that of the respective coefficient on the right-hand side of the equation. Let the estimated value of m^d from this cointegrating vector be $\hat{m}^d$. That is,

$$\hat{m}^d_t = \hat{a}_0 + \hat{a}_R R_t + \hat{a}_y y_t$$

The error-correction model is then specified as:

$$Om^d_t = \alpha z_t + \beta(m^d_{t-1} - \hat{m}^d_{t-1}) + \gamma O x_t + \eta_t \qquad (31)$$

where $(m^d_{t-1} - \hat{m}^d_{t-1})$ is the *lagged* error and *z* is a vector which includes the constant term and any I(0) variables. Since *x* is the vector of the independent variables which are I(1) and included in the cointegrating vector, $Ox_t = (x_t - x_{t-1})$ is I(0). Under our assumptions on the money demand function, *x* would include *R* and *y*, which were assumed to be I(1) and are in the theoretical specification of the demand function. Since there are no other independent variables in this function, *z* would consist only of the constant term.

But if only *m* and *y* were I(1) while *R* was I(0), the cointegration would be appropriate only over *m* and *y*. If the estimated cointegrating vector met the theoretical restrictions for the money-demand function and therefore was accepted as the long-run money demand function, the error-correction equation (31) would specify *z* by a constant term and *R*, while *x* would be specified by the single variable *y*.[18]

To conclude this section, given that the data on the money stock and income – and possibly on other variables in the money demand function – are almost always at least I(1), it is inappropriate to use the standard least-squares regression methods. This has led to the popularity of the cointegration–ECM procedure for the estimation of money demand functions. An appealing feature of this procedure is the separation of the long-run money demand function from its dynamic short-run form in a simultaneous econometric estimation of the two. One defect of the Johansen procedure arises if one or more of the variables are I(0) but have a structural break, which makes their series appear to be I(1).[19]

Some cointegration studies of the money-demand function

We examine a few studies that used the cointegration–ECM for their findings. Among these, Baba *et al.* (1992) considered the standard money-demand equation, with only the interest rate and income as the explanatory variables to be misspecified for several reasons. They claimed that these variables suffer from the omission of the inflation rate, inadequate inclusion of the yield on money itself, inadequate adjustment for financial innovation in the yields on alternative assets, exclusion of the risk and yield on long-term assets and, finally, improper dynamic specification. On the last item, they considered the partial adjustment model or the usual corrections made for serial correlation, such as the Cochrane–Orcutt technique, to be unacceptable for various reasons.

Baba *et al.* therefore estimated a more elaborate M1 demand function using the cointegration–ECM technique for the USA for 1960–88. They reported finding a stable cointegrating M1 demand function consistent with theory. Further, their finding was that the short-run money demand dynamics were adequately captured by the error-correction specification. They found a significant impact of inflation, apart from those of interest rates, on M1 demand. The inclusion of a long-term bond yield, adjusted for risk, was also significant and important for explaining the changes in velocity. However, their variable for the yield on alternative assets was a construct which included adjustments for the changing availability of financial instruments and the time required in the learning process for these instruments to be fully adopted. They concluded that if the yield data is not suitably adjusted for these factors, the mere inclusion in the estimated equations of the own-interest rates on financial assets will lead to the rejection of parameter constancy and stability.

These findings of the Baba *et al.* study point to the usefulness of the cointegration–ECM technique and the need to specify properly the variables in the money-demand function. They also stressed that financial innovation had been significant. This leads to instability of the estimated function unless the financial innovation and its pace are properly captured by the data. Unfortunately, the method that works best for capturing this in one study for a given country and given period does not often do equally well over other periods or for other countries, so that the methods for capturing innovations remain varied and somewhat eclectic.

Miller (1991) used the demand for nominal money balances as a function of real income, the nominal interest rate and the price level. His specification for money included M1, M1A, M2 and M3. The alternatives used for the interest rate were the four to six-month commercial paper rate and the dividend/price ratio. The Engle–Granger cointegration–ECM technique was used on the US quarterly data for 1959–87. Of the various monetary aggregates, only M2 was cointegrated with the other variables; none of the other ones were cointegrated.

Hafer and Jansen (1991) used the Johansen procedure for US quarterly data for 1915–88 and for 1953–88. In one part of their study, their variables were real money balances, real income and the commercial paper rate, which is a short-term interest rate. They found a cointegrating vector for M1 for 1915–88, though not for 1953–88, and found such vectors for M2 for both periods. For M1 for 1915–88, the long-run income elasticity was 0.89 and the long-run interest-rate elasticity was –0.36. For M2, the former was a plausible 1.08 for 1915–88 and a plausible 1.06 for 1953–88. The long-run interest-rate elasticity for M2 was –0.12 for 1915–88 and –0.03 for 1953–88, with both estimates being statistically significant. These estimates, especially the latter one, are much lower than the corresponding estimated elasticities in the range –0.15 to –0.5 in many other studies.

When Hafer and Jansen replaced the commercial paper rate by the corporate bond rate – a long-term rate – there was still no cointegrating vector for M1 for 1953–88. There was also none for M2 for 1915–88, but there was one for 1953–88. The income elasticity for the latter was 1.13 and the interest rate was –0.09. Overall, the authors concluded in favor of using M2 over M1 in a long-term relationship with income and interest rates.

Among other studies, Miyao (1996) used M2 for his money variable and estimated a variety of linear functions involving income, an interest rate and the price level. His sample periods for US quarterly data were 1959–88, 1959–90 and 1959–93. For the earlier periods, there were mixed results suggesting both cointegration and no cointegration, while there was no cointegrating vector at all for 1959–93. The author concluded that there were shifts in the data structure in the 1990s, so that an error-correction model was not appropriate for that decade. Further, his conclusion was that a stationary relationship between M2 and output disappeared in the 1990s, so that M2 was no longer a reliable indicator or target for policy purposes.

As pointed out earlier, innovations have shifted the money-demand function over time. Further, cointegration analysis requires long runs of data. To accommodate these, Haug (2006) uses cointegration techniques with unknown shift points to study the demand for M0, M1, M2 and related money measures for Canada covering several periods, the longest one being 1972–97. Among other criteria for acceptance of findings, Haug uses cointegration rank stability. This study also introduces variables (such as the ratio of currency to the money supply, velocity, and per capita permanent income) that reflect institutional and structural change. The findings, using the long-term interest rate, do show one cointegrating vector for the demand for M1, irrespective of the data time span.[20]

As against studies using M1 or M2 as the preferred monetary aggregate for the Canada and USA, the European Central Bank uses M3 as its preferred monetary aggregate. Coenen and Vega (2001) use cointegration and error-correction analysis to estimate the demand for M3 for the Euro area for the period 1980:Q4 to 1998:Q4. They find a stable long-run demand function for real M3. Their explanatory variables included, in addition to real GDP, short-term and long-term interest rates and the inflation rate. Their estimated long-run income elasticity is 1.13, which they interpret as incorporating wealth effects on money demand.

These differing results clearly indicate that the evidence for recent decades on the cointegration of the variables in the money demand function is not unambiguous or robust for the United States. Similar findings have been reported for the UK (see Cuthbertson, 1991, for a review of some of these) and Canada. While the existence of such a vector cannot be rejected for some form of the monetary aggregate and for some definitions of the independent variables, such a finding is dependent on particular definitions, particular periods and particular cointegration techniques (for instance, see Haug, 2006). Part of the reason for the conflicting findings is the sensitivity of the Johansen cointegrating procedures to the sample size and its poor finite sample properties. But, from the perspective of economic theory, the problem can also stem from numerous shifts in the money demand function due to innovations of various types in recent decades. These shifts imply that there is no stable long-run money-demand relationship over this period. Therefore, the cointegration techniques will not yield the appropriate cointegrating vector, unless the impact of the innovations is somehow first adequately captured in the measurement of the variables, as in the Baba, Hendry and Starr study cited above, and perhaps not even then, since the innovations have been of numerous types and their collective combination has itself been changing.

Causality

Since the ECM incorporates lags of the explanatory and other exogenous variables on the right-hand side, its estimates are often used to determine the direction of Granger causality. The criteria for judging one-way versus two-way Granger causality were specified in Chapter 7.

An illustration: money demand elasticities in a period of innovation

Table 9.1 provides an illustration of the estimates of money demand with a lagged dependent variable and is based on Goldfeld and Sichel (1990). Part of this table is based on Fair (1987), who presented the estimates of money demand for 27 countries.

Income elasticities in Table 9.1

In Table 9.1, the coefficient of the income variable *y* is the impact elasticity for the quarter and lies in the range 0.039 to 0.118. The long-run elasticity is obtained by dividing the impact elasticity by one minus the coefficient of the lagged dependent variable m_{1}. The computation of long-run elasticity becomes extremely sensitive to small changes as this coefficient approaches one. In fact, if this coefficient is one or over one, the partial adjustment model leads to a misspecification in its adjustment mechanism. This is clearly so for the USA for 1952:3–1979:3, and almost so for 1974:2–1986:4. The estimates for these periods therefore cannot be relied upon, as a look at the long-run income and interest-rate elasticities clearly shows.

Table 9.1 Estimates of money demand

Country	*Sample period*	y	R_1	R_2	π	m_{-1}	*Long-run elasticities*	
							Income	*Interest*[c]
USA[a]	1952:3–1974:1	0.131	−0.016	−0.030	−0.771	0.788	0.62	−0.075
	1952:3–1979:3	0.039	−0.013	−0.002	−0.889	1.007	−5.57	1.857
Canada[b]	1974:2–1986:4	0.044	−0.018	0.100	−0.823	0.997	14.67	−6
	1962:1–1985:4	0.071	−0.004		−1.66	0.94	1.18	−0.067
UK[b]	1958:1–1986:1	0.118	−0.005		−0.69	0.44	0.21	−0.009

Source: Goldfeld and Sichel (1990), Tables 8.1 and 8.5, of which Table 8.5 is based on Fair (1987).

Notes

[a] All variables are in logs, except for the inflation rate $\pi = (\ln(P_t/P_{t-1}))$. The dependent variable — is real money balances *m*, measured by the real value of M1, and *y* is real GNP. R_1 is the commercial paper rate and R_2 is the passbook savings rate at commercial banks.

[b] All variables are in logs except the interest rate R_1 which is in levels. The dependent variable *m* is real balances per capita and the scale variable is income per capita. r_1 is a short-term rate. The reported estimates are taken from Goldfeld and Sichel (1990), Table 8.5, calculated by them from Fair (1987).

[c] Based on the coefficient of R_1.

Further, only two of the long-run income elasticities in Table 9.1 are plausible. These are 0.62 for the USA for 1952:3–1974:1 and 1.18 for Canada. The estimates of this elasticity for the other two periods for the USA are implausible and, as already argued in the preceding paragraph, the estimated equation as a whole for these periods is highly suspect. Further, Goldfeld and Sichel show that the estimates perform well in simulations only for the first period and that the money demand function shifts sufficiently after 1974 to lead to a breakdown of its estimation in the conventional form used for this table.

Interest-rate elasticities in Table 9.1

From column 4 of Table 9.1, the impact (first quarter) interest-rate elasticities are 0.004 for Canada, 0.005 for UK and 0.016 for the first period for the USA, ignoring the latter two periods for this country. The corresponding long-run interest elasticities are 0.066 for Canada, 0.009 for UK and 0.075 for the USA. For comparison, for Canada for 1956:1–1978:4, Poloz (1980) had reported for M1 the impact and long-run interest rate elasticities of 0.054 and 0.18 respectively. His estimates of the corresponding income elasticities were 0.22 and 0.73. These are somewhat different from those reported in Table 9.1 for Canada, and indicate that one should think in terms of the plausible ranges rather than precise magnitudes for elasticities.

This table also shows significant impact elasticities with respect to the inflation rate, which are in fact higher than the interest-rate elasticities. Since the coefficient of the lagged dependent variable equals $(1 - \lambda)$, the adjustment during the first quarter was only 0.212 for the USA for the first period, 0.06 for Canada and 0.56 for the UK. We have already commented on the instability of the money demand function in the latter two periods for the USA. Fair found instability for 13 out of the 17 countries in his sample.

As discussed in Chapter 4, the Baumol–Tobin inventory model of transactions money demand implies that, at relatively low interest rates relative to brokerage costs, it may not be optimal for economic agents to hold bonds for transactions purposes, whereas doing so would become optimal at higher interest rates, so that the interest elasticity of the transactions demand would vary between –1/2 and 1. Therefore, as Mulligan and Sala-i-Martin (2000) argued, the interest elasticity would be non-linear. Their findings confirm that the interest elasticity of money demand is low at low interest rates.

Innovations and the search for a stable money-demand function

Financial innovation is a frequent occurrence in the economy. Some types of innovation change the liquidity characteristics of the existing assets or represent the creation of new assets. Other types of innovation are in the payments and banking technologies. Some of the innovations could also be due to the attempt of the financial industry to get around financial regulations. Another is the introduction of new techniques of financial management by firms, households and financial institutions. All of these have occurred during the last three decades, probably collectively at a faster pace than in earlier decades.

Among the new types of assets, in the USA, interest-bearing checking accounts were first introduced as NOW (negotiable orders of withdrawal) and then as super-NOW accounts in the late 1970s and early 1980s. Commercial banks began to issue small certificates of deposit in the 1960s and money-market mutual funds in the late 1970s. These were outside the traditional definition of M1. In the UK, commercial banks and building societies introduced checkable interest-bearing accounts in the 1980s. In each case, there was a learning period for the public and shifts in the money demand function were evident over many years.

If the innovations merely change the constant term or the coefficients of the independent variables in the money demand function, they can be relatively easy to capture in estimation through period splitting or the use of dummy constant and interactive variables. However, some of the resulting shifts of the money demand function are much more difficult to capture or cannot be captured, and the researcher ends up with the judgment that the money demand function has become unstable.

The desperate search for a stable money demand function

The last three decades have seen a remarkable number of innovations in the monetary sphere. These have resulted in a breakdown of the estimated money demand functions and a large number of innovations by researchers in their estimating equations and techniques. The attempts to find a stable demand function have included changes in the monetary aggregate used as the dependent variable (M1, M2, M3, or their Divisia counterparts). Other attempts have centered around variations in the arguments of the function. These included the use of current income, permanent income, wage income or property income, etc., for the scale variable, and the use of short interest rates, long interest rates, the rate of inflation or a composite index of interest rates, etc., for the interest rate variable.

Still other attempts changed the form of the estimating equation from linear to log-linear and semi-log-linear, or switched to non-linear functions or tried ones with stochastic coefficients, or used transcendental functions. Some other attempts focused on the proper specification of the dynamic adjustment of the actual to desired money balances. The econometric techniques have included the classical regression techniques and cointegration–error-correction models, among others.

This prolific variety of attempts and deviations from the standard money demand equation almost gives one the impression of a field dominated by data mining and the *ad hoc* constructions of a profession desperate to find a stable money demand function to back its theory. While this may sound a rather harsh assessment, it does serve as a reminder of the severe difficulties in finding a stable money demand function during the ongoing innovations of the recent decades.

For the USA, there appears to have been a downward shift in the demand function during the 1970s and an upward shift during the 1980s. In these decades, as in the 1990s, actual money holdings deviated remarkably from the predictions of most estimated money demand models. In terms of velocity, the velocity of M1 increased in the 1970s and decreased in the 1980s in a manner not predicted by these models.

Conclusions

Empirical findings generally confirm the homogeneity of degree zero of the demand for real balances with respect to the price level – and the consequent homogeneity of degree one of the demand for nominal balances – as discussed in Chapter 3. The income elasticity of real M1 with respect to real income has been established as being less than one, even in the long run, though some studies show the income elasticity for real M2 to be even slightly larger than unity. The latter is particularly so for developing economies, in which the bond and stock markets are not well developed, so that increases in savings are mostly held in savings deposits. Real balances do depend on interest rates, with a short-term rate being usually used in the estimation of M1 demand and a longer term one being used for the estimation of M2 demand. The estimated interest-rate elasticities usually fall in the range from _0.15 to _0.50. In the LDCs, the rate of inflation typically performs better in estimation than the rate of interest and is often used in lieu of the latter, with somewhat similar elasticities. While currency substitution is a theoretical possibility and some studies do confirm its existence for their data sets, empirical studies have not always found it to be so significant that the elimination of the return on foreign currencies from the money demand function leads to much worse results. Most money demand functions are, therefore, estimated without this variable. Not much support has been found for the liquidity trap and it is now hardly ever investigated or even mentioned in empirical studies.

The velocity of circulation of M1 is not a constant in either the short or the long run. In the short run, its annual variation is quite significant even in stable economies without political and economic panics. It is about 3 percent to 4 percent for the USA, but can be much higher in less stable economies. Since the income elasticity of M1 is likely to be less than one in the long run, the long-run expectation for its velocity is that it will increase.

Innovations in the financial sector and in the usage of money by non-financial economic agents in the economy have been very rapid in the last three decades, so that the money demand functions estimated with data including this period are often not stable. Further, it is even more rare to find the estimated functions for both narrow and wide definitions of money to be stable for a given country over a given period.

For the open economy, the existence of extensive CS could cause the monetary authority to lose control of the domestic money supply and increase the volatility of exchange rates under a floating exchange rate regime. It would increase the speculative pressures under fixed exchange rates. A common finding in estimations of the open-economy domestic money demand function is that the expected change in the exchange rate – which is the proxy on the return on holding foreign money relative to that on domestic money – is not significant in explaining domestic money demand. Such a finding has led to the conclusion that currency substitution tends to be extremely low, even in countries like Canada, in which the public often holds US dollars in currency or in US dollar bank deposit accounts. However, many studies also show that the return on foreign bonds is a significant positive determinant of domestic money demand. This could provide indirect evidence on CS: if foreign money balances provide monetary services as a medium of payments in the domestic economy, the decrease in their holdings due to an increase in the return on foreign bonds, has to be compensated by an increase in domestic money balances in order to maintain the desired holdings of all media of payments. This effect relies on the substitution between the domestic and foreign monies in their medium-of-payments role, while relying upon substitution between the foreign money and foreign bonds in portfolio allocation.

Several of the variables crucial to money demand estimation are not stationary. This is especially likely to be so for the monetary aggregates themselves, as well as for the income and wealth variables. It may or may not also be so for the interest rates in the particular data set. Consequently, the classical regression techniques do not yield unbiased and consistent coefficients. Cointegration analysis is an appropriate procedure in this case and has become quite common in recent years for estimating money demand functions. Its combination with error-correction modeling has the further advantage that the estimation yields both the long-run and the short-run demand functions.

Cointegration procedures represent an attempt to capture the long-run equilibrium relationship and there should exist a cointegrating vector if such a relationship is stable over the sample period. However, when long-run relationships are shifting due to innovations and the impact of the innovations has not been eliminated or somehow captured in the definition of the variables or the procedure used, the sample data would not incorporate a stable long-run relationship.

Money demand studies using cointegration techniques for data over the last few decades have provided a mixed bag of evidence about the existence of a cointegrating vector between money, income, interest rates and prices. The finding of such a vector has often been culled from the data by using different definitions of money, different interest rates and different periods. The last few decades have seen a mixed bag of very significant innovations related to money demand, so that the long-run money demand function must have been shifting. Consequently, cointegration studies, like earlier studies using the standard regression techniques, have not provided convincing evidence of the existence of a stable long-run money-demand function for the last few decades for Britain, Canada and the USA.

In studies where acceptable cointegration vectors have been established, an error-correction model has usually also been estimated. As expected, these studies show for quarterly data that the impact elasticities are relatively quite small and much smaller than the long-run elasticities, indicating that adjustments of money demand to changes in the independent variables take at least several quarters.

We have not differentiated between the demand functions estimated for the different segments of the economy, such as households, firms and financial institutions. There are numerous studies on these and the interested reader is encouraged to explore them. There is a significant difference between the demand for money by households and that by firms, especially large ones. In general, the former tends to be relatively more predictable than the latter.

Appendix

The ARDL model and its cointegration and ECM forms

As explained in Chapter 8, the regressors in an autoregressive distributed lag (ARDL) model include the lagged values of the dependent variable and the current and lagged values of the explanatory variables. Its estimating equation with p lagged values of the dependent variable and q_j, $j = 1, 2, \ldots, n$, values of the n explanatory variables is designated as an ARDL($p, q_1, \ldots, q_n$) and has the form:

$$\beta(L,p)y_t = \beta_0 x_0 + \sum_{j=1}^{n} \beta_j(L,\ q)x_{jt} + \mu_t \tag{32}$$

where L is the lag operator such that $L^i y_i = y_{t-i}$, x_0 is a constant and (L, p) and (L, q) are the lag polynomials:

$$\alpha(L,p) = 1 - \alpha_1 L^1 - \alpha_2 L^2 - \alpha_p L^p \tag{33}$$

$$\beta(L,q) = 1 - \beta_1 L^1 - \beta_2 L^2 - \alpha_q L^q \tag{33'}$$

In the long run, $y_t = y_{t-1} = \cdots = y_{t-p}$ and $x_{jt} = x_{jt-1} = \cdots = x_{jt-q}$, so that $L = 1$, $\alpha(1,p) =$

$(1 - \alpha_1 - \alpha_2 - \alpha_p)$ and $\beta(1,q) = (1 - \beta_1 - \beta_2 - \beta_q)$ and the long-run relationship becomes:

$$y_t = \beta'_0 + \sum_{j=1}^{n} \beta'_j x'_{jt} + v_t \tag{34}$$

where $\alpha'_0 = \alpha_0/(\alpha(1,p)$, $\beta_j = \beta'_j(1,q)/(\alpha(1,p)$, $v_j = \mu_j/(\alpha(1,p)$. The error-correction equation of this ARDL model is:

$$\Delta y_t = \Delta\beta'_0 - \sum_{i=2}^{p} \alpha'_i \Delta y_{t-i} + \sum_{j=1}^{n} \beta'_{j0}\Delta x_{it} - \sum_{j=1}^{n}\sum_{i=2}^{q} \beta'_{i,t-j}\Delta x_{j,t-i} - \alpha(1,p)\mathrm{ECM}_{t-1} + \eta_t \tag{35}$$

where:

$$\mathrm{ECM}_{t-1} = y_{t-1} - \hat{\beta} - \sum_{j=1}^{n} \beta_j x_{jt-1} \tag{36}$$

$j=1$

$\alpha(1, p)$ measures the speed of adjustment.

An illustration: a simple ARDL model

The simplest case of an ARDL model has only one explanatory variable x_1 and one-period lags, so that it is ARDL(1, 1). The estimation equation for this case is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \tag{37}$$

where μ is white noise. The long-run relation between y and x_1 for this equation is obtained by setting $y_t = y_{t-1}$ and $x_t = x_{t-1}$, so that the long-run equation is:

$$y = \alpha_0/(1-\alpha_1) + \{(\beta_0 + \beta_1)/(1-\alpha_1)\}x_{1t} + \{1/(1-\alpha_1)\}\mu_t \tag{38}$$

where $(\beta_0 + \beta_1)/(1 - \alpha_1)$ provides the long-run relationship between y and x. Further, in (38), replacing y_t by $(y_{t-1} + Oy_t)$ and x_{1t} by $(x_{1t-1} + Ox_{1t})$, we get:

$$Oy_t = \alpha_0 - (1-\alpha_1)y_{t-1} + \beta_0 Ox_{1t} + (\beta_0 + \beta_1)x_{1t-1} + v_t \tag{39}$$

which is the short-run ECM representation of (37).

Further, from (37), we have:

$$\begin{aligned} y_t - \alpha_1 y_{t-1} &= \alpha_0 + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \\ (1 - \alpha_1 L)y_t &= \alpha_0 + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \end{aligned} \tag{40}$$

Expanding $\{1/(1 - \alpha_1 L)\}$, we have:

$$1/(1 - \alpha_1 L) = (1 + \alpha_1 + \alpha_1^2 + \cdots)$$

Hence, (40) yields:

$$y_t = (1 + \alpha_1 + \alpha_1^2 + \cdots)\alpha_0 + (1 + \alpha_1 + \alpha_1^2 + \cdots)(\beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t) \tag{41}$$

which is another way of stating (37). In this context, $(1 + \alpha_1 + \alpha_1^2 + \cdots)(\beta_0 + \beta_1) = (\beta_0 + \beta_1)/(1 - \alpha_1))$ provides the long-run relationship between *y* and *x*.

Part IV

Monetary policy and central banking

10 Money supply, interest rates and the operating targets of monetary policy

Money supply and interest rates

This is the first of three interrelated chapters on monetary policy and central banking. It starts by examining the goals and operating targets of monetary policy. The two major operating targets of monetary policy are the money supply and the interest rate.

This chapter then focuses on the determination of the money supply. While macroeconomic models tend to simplify by assuming that the money supply is exogenously determined, the private sector in the form of the banks, households and firms also influences the money supply.

Key concepts introduced in this chapter

- Targeting inflation or its deviation from a desired inflation rate
- Targeting output and unemployment
- Interest rate as an operating target
- Monetary base
- Currency ratio
- Demand deposit ratio
- Free reserves
- Excess reserves
- Required reserves
- Discount/bank rate
- Mechanical theories of the money supply
- Behavioral theories of the money supply

This is one of three chapters on some of the central issues of monetary policy. It starts with the relationships among the goals, intermediate targets and operating targets of monetary policy and examines the theoretical justification as well as the implications of adopting different targets. It then considers the issue of whether the central bank should use the money supply or the interest rate as its major monetary policy instrument. It then narrows its focus to the determination of the money supply in the economy, so as to complement the extensive treatment of money demand in the preceding chapters.

Sections 10.1 and 10.2 present the links between the goals and targets of monetary policy. Sections 10.3 to 10.5 examine the main operating targets of monetary policy commonly used by central banks and their justification from macroeconomic analysis.[1] Sections 10.6 to 10.8 present the determination of the money supply. Section 10.9 covers the application of cointegration analysis and error-correction modeling to money supply. Section 10.10 considers the central bank's choice between the monetary base and the interest rate as alternative operating targets.

Stylized facts on the goals and operating targets of monetary policy

The stylized facts on monetary policy depend on the behavior of the central bank and the structure of the economy. Among these facts are:

1 The central bank has more than one goal. Among its goal variables are output and its growth rate, unemployment, inflation, etc. Currently many central banks focus on reducing the deviation of output from its full-employment level and of inflation from a target level, with a trade-off between them, as in a Taylor rule.
2 The target inflation rate for many central banks now is a low inflation rate, often in a range of 1 percent to 3 percent.
3 The operating target of monetary policy can be a monetary aggregate or an interest rate. A monetary aggregate was selected for this purpose in some past periods and is still in use by some central banks. Currently, many central banks in the developed economies focus on an interest rate as their primary operating target.
4 The central bank does not control the money supply directly but has to use its instruments, such as the monetary base, for indirectly controlling the money supply.

Goals, targets and instruments of monetary policy

The eventual purpose of monetary policy is to achieve certain national goals. These have historically included full employment (or a low unemployment rate), full-employment output (or a high output growth rate), a stable price level (or a low inflation rate), a stable exchange rate (or a desirable balance of payments position), etc. These variables are simply referred to as "*goals*" or as "ultimate goals" of monetary policy. However, the central bank cannot achieve these goals directly by its monetary policy *instruments*, which are variables that it can operate on directly. Among the instruments available to the central bank are open-market operations and changes in its discount/bank rate at which it lends to commercial banks and other bodies. These determine the economy's monetary base. In many countries, the central bank can also change the required reserves (i.e. the minimum reserves the commercial banks must hold against the public's deposits with them), which changes the "monetary base multiplier" (i.e. the money supply per dollar of the monetary base). These measures serve to change the money supply in the economy. Another monetary policy instrument is the overnight loan rate (called the federal funds rate in the USA) in the market for reserves, whose operation induces change in various interest rates in the economy. The next chapter provides further information on the goals and instruments of monetary policy.

Besides the concepts of goals and instruments, other concepts relevant to monetary policy are those of targets, operating targets and guides. We can broadly define a *target* variable as one whose value the policy maker wants to change.[2] An *operating target* variable is one on which the central bank can directly or almost directly operate through the instruments at its disposal. A *guide* is a variable that provides information on the current and future state of the economy.

Between the goals and instruments of monetary policy lie layers of intervening variables. For example, suppose the central bank wants to reduce the inflation rate. To do so, it needs to reduce aggregate demand in the economy. The reduction in aggregate demand usually requires a reduction in investment and/or consumption, which requires an increase in market interest rates. Depending on the analysis, discussion or author, these intervening variables can be referred to as intermediate targets, operating targets or even as instruments. Since a target variable is one whose value the central bank seeks to influence or control by the use of the tools at its disposal, any of the intervening variables between the goals and instruments can be referred to as a target variable. In the preceding example, aggregate demand is an intermediate variable or target, which the central bank wants to alter by using the intermediate targets of the money supply and/or interest rates which, in turn, can be altered by changes in the monetary base and the discount rate. Note that the word “target” can also be used to indicate a desirable value of a goal (e.g. inflation) or of an intermediate variable (e.g. the money supply and market interest rates).

Given the preceding discussion, Table 10.1 provides a rough classification of monetary policy instruments, operating targets, intermediate targets and goals.

While Table 10.1 provides some guidance on the roles and sequence of the various monetary policy variables, there is no hard and fast rule for its classification. The central bank uses its tools to hit its operating targets, with the intention of manipulating the intermediate targets, which are the final ones of the financial system, in order to achieve its goals. Note that lags enter at each stage of this process, and both the individual lag and the overall lag tend to vary. Further, the duration of the lags and the final impact are not usually totally predictable.

Table 10.1 Monetary policy tools, target and goals

Policy instruments	*Operating targets*	*Intermediate targets*	*Goals*
Open-market operations Discount rate	Short-term interest rates Reserve aggregates	Monetary aggregates (M1, M2, etc.)	Low unemployment rate Low inflation rate
Reserve requirements	(monetary base, reserve, nonborrowed reserves, etc.)	Interest rates (short and long term)	Financial market stability Exchange rates
		Aggregate demand	

Relationship between goals, targets and instruments, and difficulties in the pursuit of monetary policy

Several issues arise in the selection and use of goals, intermediate variables and operating targets or instruments by the monetary authorities. Among these are:

1 Are the relationships between the ultimate goal variables, intermediate variables and operating targets stable and predictable?

2 Can the central bank achieve the desired levels of the operating targets through the instruments at its disposal?

3 What are the lags in these relationships, and, if they are long, can the future course of the economy be reasonably well predicted?

To illustrate these points, let the relevant relationships be:

$$y = f(x; \Psi) \tag{1}$$

$$x = g(z; \theta) \tag{2}$$

where:

y = (ultimate) goal variable
x = intermediate target
z = policy instrument or operating target
Ψ, θ = sets of exogenous variables

The above equations imply that:

$$y = h(z; \Psi, \theta) \tag{3}$$

so that z can be used to achieve a desired value of y. However, this can be done reliably only if the functional forms f and g are known and these are stable univalued functions.[3] In practice, given the complex structure of the real-world economies, as well as the existence of uncertainty and lags in the actual relationships, the precise forms of f, g and h are often only imperfectly known at the time the decisions are made. Further, the coefficients in these relationships may be subject to stochastic changes. In addition, given the lags in the economy, the policy maker also needs to predict the future values of the coefficients and the exogenous variables – again, usually an imprecise art.

Hence, the precision and clarity implied by (3) for the formulation of monetary policy and its effects is misleading. In many, if not most instances, the impact of a change in most of the potential operating variables on the ultimate goals is likely to be imprecise, difficult to predict and/or unstable. This makes the formulation of monetary policy an art rather than a science and cautions against attempts to use monetary policy as a precise control mechanism for “fine-tuning” the goals of such policy.

Another common problem with most target variables is that they are endogenous and their values depend on both demand and supply factors, so that the exogenous shocks to them could come from either demand or supply shifts. The policy maker may want to offset the effect of changes in some of these factors but not in all cases, so that it needs to know the source of such changes before formulating its policy.

Targets of monetary policy

The two main *operating targets* usually suggested for monetary policy are:

- monetary aggregates;
- interest rates.

The two main *targets* of monetary policy highlighted in the recent literature are:

- inflation rate (or the price level),[4] or its deviation from a desired value;
- output, or its deviation from the full-employment level.

There are also other variables that are sometimes used or proposed as the intermediate targets of monetary policy. Among these is aggregate demand (or nominal national income) and, in the case of relatively open economies, the exchange rate or the balance of payments. For the sake of brevity, this chapter discusses only the relative merits and demerits of monetary aggregates and the interest rate as the chief operating target or instruments. It also presents some discussion of the price level and the inflation rate, and the output gap, as the targets of monetary policy.

Monetary aggregates versus interest rates as operating targets

This section relies upon students' prior knowledge of the IS–LM macroeconomic model (otherwise, see Chapter 13) to distinguish between the relative merits of using the money supply versus interest rates as the operating target of monetary policy. The choice between monetary aggregates and the interest rate depends critically upon the policy objective of the central bank and the structure of the economy. The following analysis, adapted[5] from that in Poole (1970), takes this objective to be control of aggregate demand,[6] since the central bank can only influence output and inflation, which are its final goal variables, through manipulation of aggregate demand. It further assumes that the structure of the economy can be represented by the IS–LM analysis and diagram. This diagram has aggregate real demand *y* on its horizontal axis and the real interest rate *r* on its vertical axis. The commodity market equilibrium is shown by the IS curve and the money market equilibrium is shown by the LM curve. Their intersection determines real aggregate demand at the existing price level.

Therefore, the choice between the monetary instruments hinges on the question: which instrument provides better control over aggregate demand in the IS–LM framework? Our analysis implicitly assumes the Fisher equation for perfect capital markets and an expected inflation rate of zero, so that the nominal interest rate R is identical with the real interest rate r.

Since the IS–LM analysis has not yet been mathematically covered in this book, this chapter presents only the diagrammatic analyses of monetary versus interest rate targeting. Its mathematical version is presented in Chapter 13, which could be read at this point.

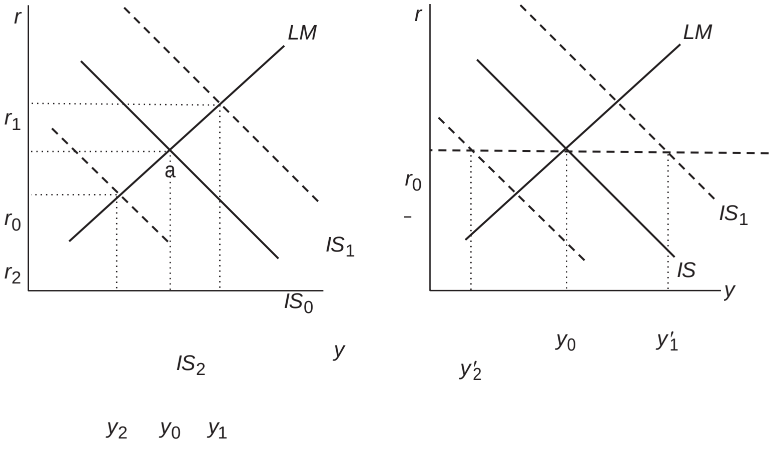
Diagrammatic analysis of the choice of the operating target of monetary policy

Shocks arising from the commodity market

The IS equation and curve encompass the various components of expenditures, such as consumption, investment, exports, fiscal deficits, etc., in the economy (see Chapter 13). Several of these are volatile, with investment often being the most volatile component of expenditures. Shifts in any of these components shift the IS curve in the IS–LM diagram.

Our analysis starts with the initial equilibrium shown by point a, with coordinates (r_0, y_0), in Figure 10.1a. Assume that the central bank targets the money supply and holds it constant through open market operations or by the use of some other instruments. Shocks to the IS curve[7] would then change both r and y. To illustrate, if a positive shock shifts the IS curve from IS_0 to IS_1, aggregate demand will increase from y_0 to y_1 and the interest rate rise from r_0 to r_1. Similarly, a negative shock, occurring, say, in the following period, which shifts the IS curve to IS_2, will lower aggregate demand to y_2 and the interest rate to r_2.

Compare this result with the impact of the same shock if the interest rate had been targeted. This is shown in Figure 10.1b, where the interest rate is assumed to be held fixed by the authorities at the target rate $\underline{r}_0$, where the underline indicates that it is exogenously set by the central bank. The shifts in the IS curve, first to IS_1 and then to IS_2, will produce movements



(a) (b)

Figure 10.1

in aggregate demand, first to y_1^J and then to y_2^J. This fluctuation between y_1^J and y_2^J is clearly greater than between y_1 and y_2 in Figure 10.1a, so that targeting the interest rate produces greater fluctuations in aggregate demand than money supply targeting if the exogenous shocks emanate from the commodity market. Note that such shocks do not produce changes in the interest rate, since that is being held constant through monetary policy.

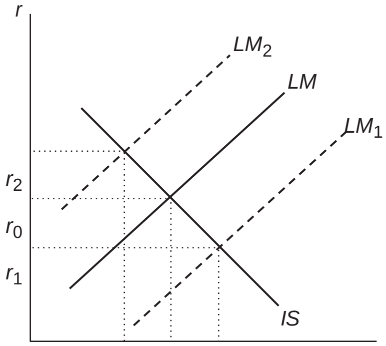
Shocks arising from the money market

Now assume that the exogenous shocks arise only in the money market while there are no shocks in the commodity market, so that the IS curve does not shift. Such exogenous shocks in the money market can be to either money demand or money supply, and shift the LM curve.

Money supply targeting would stabilize the money supply,[8] so that disturbances to it do not have to be considered, but not the money demand. Now suppose that money demand decreases. Given the targeted money supply, the decrease in the money demand will shift the LM curve in Figure 10.2 to the right to LM$_1$ and increase aggregate demand from y_0 to y_1. Assume that the next period's shock to the money demand increases it and shifts the LM curve to LM$_2$, so that aggregate demand falls to y_2. The aggregate demand fluctuations are then from y_1 to y_2 and the interest rate fluctuations are from r_1 to r_2.

For interest rate targeting, assume that the real interest rate had been set at r_0, as shown in Figures 10.3 and 10.4. Figure 10.3 shows the initial demand curve for nominal balances as M^{d} and the initial supply curve as M^{s}, with the initial equilibrium interest rate as r_0 and the initial money stock as M_0. Now suppose that the money demand curve shifts from M_0^{d} to M_1^{d}. Since the interest rate is being maintained by the monetary authority at $\underline{r}_0$, the monetary authority will have to increase the money supplied from M_0 to M_1. The money stock therefore adjusts endogenously through an accommodative monetary policy to the changes in money demand.

In the IS–LM Figure 10.4, a reduction in the money demand would shift the LM curve to the right from LM$_0$ to LM$_1$. However, given that the monetary authority maintains the



y_2 y_0 y_1 y

Figure 10.2

[3] However, monetary base targeting usually will not do so.

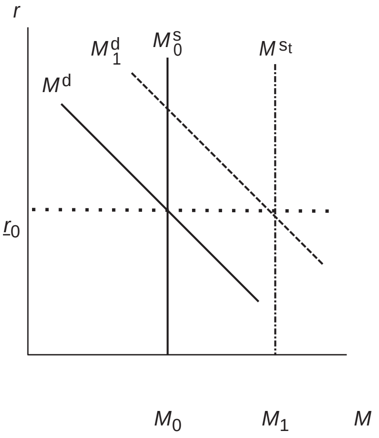


Figure 10.3

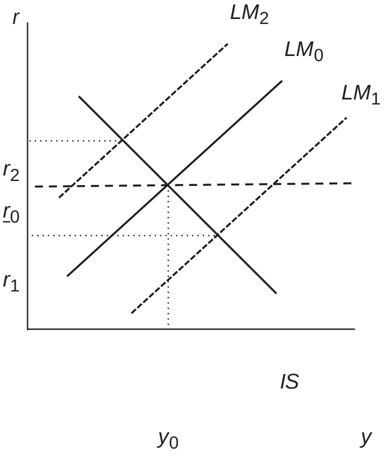


Figure 10.4

interest rate at $\underline{r}_0$, the aggregate demand y_0 in this figure will be determined by the intersection of the IS curve and a horizontal line at the target interest rate $\underline{r}_0$. This is so because the exogenous shift in the LM curve from LM_0 to LM_1 leads the central bank to undertake an accommodative money supply decrease sufficient to shift this curve back to LM_0. Hence, in spite of any exogenous changes in money demand, aggregate demand would remain at y_0 (and the interest rate at $\underline{r}_0$). Hence, comparing the implications from Figures 10.2 and 10.4, monetary targeting will allow greater fluctuations in aggregate demand and interest rates than interest-rate targeting when the exogenous shifts arise from money demand.

This conclusion poses a problem for the policy maker since both types of shocks occur in the real world. Therefore, the monetary authority has to determine the potential source of the dominant shocks to the economy before making the choice between monetary and interest rate targeting. This is not easy to determine for the future, nor need the same pattern of shocks necessarily occur over time. Further, since both types of shock do occur, each policy will reduce or eliminate the impact of some types of shocks but not of others.

While many central banks had, for a few years during the late 1970s and sometimes in the early 1980s, favored monetary targeting, the common practice currently is to set interest rates.

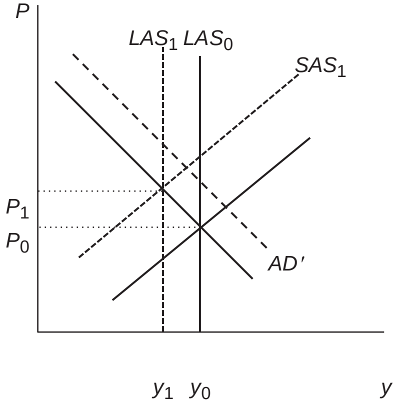


Figure 10.5

This implies, in the context of the preceding analysis, that the dominant sources of shocks are expected to be in the monetary sector.

Analysis of operating targets under a supply shock

For the analysis of operating targets under supply shocks, we start by changing the objective function from stabilization of aggregate demand y^d to stabilization of the price level P or/and real output y. Figure 10.5 shows the aggregate demand (AD) curve and the short-run (SAS) and long-run (LAS) aggregate supply curves for the economy. The initial equilibrium is at (y_0, P_0). An anticipated *negative permanent* supply shock will shift the supply curves from SAS_0 to SAS_1 and LAS_0 to LAS_1. First, consider the short-run effect of the fall in supply to SAS_1. Prices rise from P_0 to P_1, while output falls from y_0 to y_1. The rise in prices will decrease the real money supply and shift the LM curve to the left (for instance, from LM_0 to LM_2 in Figure 10.4), so that the interest rate will rise (from r_0 to r_2). Monetary targeting will leave the money supply unchanged and therefore leave the new equilibrium at y_1, P_1 and r_2.

But an interest rate target at $\underline{r}_0$ will cause the central bank to increase the money supply to prevent the interest rate from rising. This will increase aggregate demand and cause a policy-induced shift from AD to AD^J in Figure 10.5. The result will be a further increase in prices but the fall in output will be partly or wholly (depending on the induced demand increase) offset in the short run. The relevant intersection is that of SAS_1 and AD^J. Hence, in the short run, interest-rate targeting is more inflationary than monetary targeting but compensates for this by limiting the fall in output.

Now consider the long-run analysis with the shift from LAS to LAS_1. In this case, interest-rate targeting will cause a continual increase in the money supply and the price level, without any beneficial offset in terms of output or the maintenance of the interest rate at $\underline{r}_0$. Therefore, for permanent supply shocks, monetary targeting is clearly preferable in the long run, whereas interest-rate targeting involves a cumulative inflationary process.

Monetary aggregates as targets in practice

Milton Friedman and the 1970s monetarists, belonging to the St Louis school, had argued that because of the existence of both a direct and an indirect transmission mechanism from the money supply to aggregate expenditures, the money supply rather than interest rates provided better control over the economy. Partly as an outcome of this advice, most countries – including the USA, Britain and Canada – switched to the targeting of monetary aggregates after the mid-1970s (though only until the early 1980s). The monetary aggregates often suggested as targets were M1 or M2 – and M4 in Britain – though sometimes even broader targets were also considered.

Monetary aggregate targeting was predicated on the belief that the relationship between such a target and aggregate demand was stable and had a short and predictable lag. This was certainly the finding of the studies done by the St Louis school. Monetary targets were pursued in the late 1970s and early 1980s by the monetary authorities in the USA, Canada and UK. However, the functional relationships between the monetary variables and aggregate expenditures, let alone the rate of inflation, proved to be unstable, so that they had been abandoned by the 1990s in each of these countries. Among the reasons for this instability were financial innovations and changes in the payments technology occurring in recent decades.[9] In terms of experience during the late 1970s and 1980s, direct targeting of monetary aggregates increased both the level and the volatility of interest rates considerably, with the latter considered by many economists to be destabilizing for the economy. Attempts to control the monetary or reserve aggregates directly, as a way of controlling the economy, were abandoned by most central banks in the early 1980s in favor of interest-rate targets as the control variable. This is not to say that the monetary aggregates are not monitored and the changes in them not considered in formulating monetary policy. However, for most central banks, they have ceased to be the main operating targets.

Interest rates as targets in practice

Monetary policy acts through interest rates on spending, so that the interest rates are closer in the chain of influence on spending. Hence, they are more reliable and more appropriate indicators of the need for action than are the various measures of money supply and the monetary base. In line with this, in financially developed economies such as those of the USA, Canada and the UK, the central banks believe that interest rates are a major indicator of the performance of the economy and tend to use them as the preferred guide and operating target of monetary policy.[10]

There are several measures of interest rates that may be considered, with the usual selection for operating purposes being of short-term nominal, rather than long-term or real, rates of interest. Historically, the measure commonly used for this purpose used to be the Treasury bill rate. As discussed later in Chapter 11, more recently the USA, UK and Canada have used an overnight loan rate as an operating target. These countries have well-developed markets for overnight loans among financial institutions, with this market serving as the market for the excess reserves of banks. This market for reserves is known as the Federal Funds market in the United States and the overnight loan market in Canada and the UK. Such a rate reflects the commercial banks' demand and supply conditions for reserves. The central bank's policy actions on the monetary base immediately affect the commercial banks' demand and supply of reserves, thereby changing the overnight interest rate and starting a chain of reactions on other interest rates, and through these on the borrowing and lending, investment and consumer spending, etc., in the economy. A higher rate means that banks are relatively loaned up and a lower rate means that banks have relatively large free reserves, so that they can increase loans of their own volition.

Problems with the use of interest rates in managing the economy

The observed interest rates are equilibrium rates, so that changes in them could reflect either changes in demand or in supply conditions or both. Therefore, a rise in the interest rates may be due to an increase in the demand for loanable funds or a decrease in their supply, but the central bank may wish to take offsetting action in only one of these cases. For example, interest rates rise during an upturn in the business cycle. The central bank may not wish the upturn to be curbed by a decreased supply of funds but also may not wish to offset the stabilization effect of interest rates due to an increase in their demand. But changes in the equilibrium interest rates do not by themselves provide adequate information as to the causes of their rise and therefore as to the policy actions that should be undertaken. Consequently, central banks in practice supplement information on interest rates with other information on demand and supply conditions before making their policy decisions.

A problem with using interest rates as an operational target is that the central bank can determine the general level of interest rates but not equally well control the differentials among them. Examples of these differentials are the loan-deposit spread of commercial banks, and the spread between deposit rates and mortgage rates, if the latter are variable. Spreads depend upon market forces and can be quite insensitive or invariant to the central bank's discount rate. Financial intermediation in the economy is more closely a function of such differentials than of the level of interest rates, so that the ability of the central bank to influence the degree of financial intermediation through its discount rate and the overnight loan rate for reserves becomes diluted.

Among other problems is the lag in the impact of changes in the interest rate on aggregate demand in the economy. Among the reasons for such lags are the costs of adjustment of economic variables such as the capital stock and planned consumption expenditures, and the indirect income effects of changes in interest rates. There are two aspects of this lag: its length and variability. The former is often assessed at about six quarters to two years in the United States, Britain and Canada. While there is agreement that there is some variability in the length of the lag, there is no consensus on whether it is so long that changes in interest rates, intended to be stabilizing, can prove to be destabilizing. Within the lag, the *impact* effect (within the same quarter) of interest rate changes on real aggregate demand is estimated to be quite low, while the *long-run* effect is now believed to be very significant.

The actual use of interest rates for stabilization has often been found to be "too little, too late" – though this is usually a result of uncertainty about the need for and the lags in the effects of monetary policy. This results in its cautious use, no matter what operational or indicator variable is used. Given the duration of lags and the uncertainty at any time about the position of the economy in the business cycle, past experience does indicate that central banks often change the interest rates later and in smaller steps than really needed. An initial change is, therefore, often followed by many more in the same direction over several quarters.

Money supply under an interest rate target

In market economies, the use by the central bank of the interest rate as its major instrument of monetary policy does not imply that it can ignore the money supply altogether. Interest rates are determined in financial markets, so that if the central bank were to lower its interest rate and not provide the supporting required increase in the money supply, it would find that the market rates will diverge from its desired ones, so that the intended effects on expenditures will not be achieved. Hence, an interest-rate policy must be accompanied by an appropriate money supply. This topic is addressed in the macroeconomic context in Chapter 13.

The price level and inflation rate as targets

Targeting the price level

Current discussions of monetary policy often refer to inflation or price targeting as the goal of monetary policy. A stable price level or a low inflation rate is sometimes proposed as the *ultimate goal* of monetary policy. For this, it is argued that money is neutral in the long run, so that the central bank cannot change the level and path of full-employment output, nor should it attempt to do so since such an attempt will only produce inflation. Under this neutrality argument, what the central bank can do is to ensure a stable value of money, so that its target should be in terms of the price level or the rate of inflation. Further, a fairly stable price level reduces the risks in entering into long-term financial contracts and fixed real investments, and promotes the formulation and realization of optimal saving and investment, which in turn increase output and employment. By comparison, high and variable inflation rates inhibit economic growth by introducing uncertainty into long-term financial contracts and investment.

For the following analyses of the price level and the inflation rate as the monetary authorities' target, we leave aside the comparison of monetary versus interest rates as targets and focus on aggregate demand as the variable in the control of the monetary authority, and assume that it will adopt the appropriate instrument to achieve the desired level of aggregate demand. Further, since our analysis is short run, we use a positively sloping short-run aggregate supply curve rather than a vertical long-run one.

Figure 10.6 assumes that there is a positive *demand shock* such that the AD curve shifts to AD_1. If the monetary authorities stabilized prices at P_0, output would remain unchanged at y_0. To achieve this under monetary targeting, the monetary authority would pursue a compensatory decrease in the money supply or an increase in the interest rate to shift aggregate demand back to AD. Under interest-rate targeting, they would raise the interest rate to achieve the same effect. The net effect of such a monetary policy would stabilize both the price level and output in the event of exogenous shocks from the money or commodity markets.

Figure 10.7a shows the effects of a negative supply shock such that the short-run aggregate supply curve SAS shifts from SAS_0 to SAS_1. This will produce an increase in the price level from P_0 to P_1 and a decrease in output from y_0 to y_1. Since the price level is not an operational variable under the direct control of the central bank, the bank would have to achieve price stability through a reduction in aggregate demand, which requires a contraction of the money supply or a rise in interest rates such that AD is made to shift to AD^J. This will, however, decrease output from y_0 at P_0 to y_1 at P_1 due to the supply shock and then to y_1^J due to the contractionary monetary policy and its implied shift of the AD curve to AD^J. Hence, the

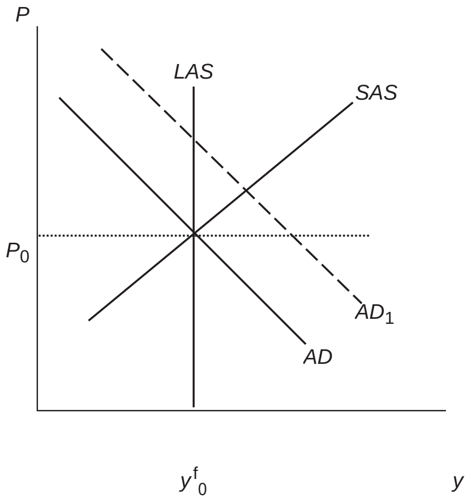


Figure 10.6

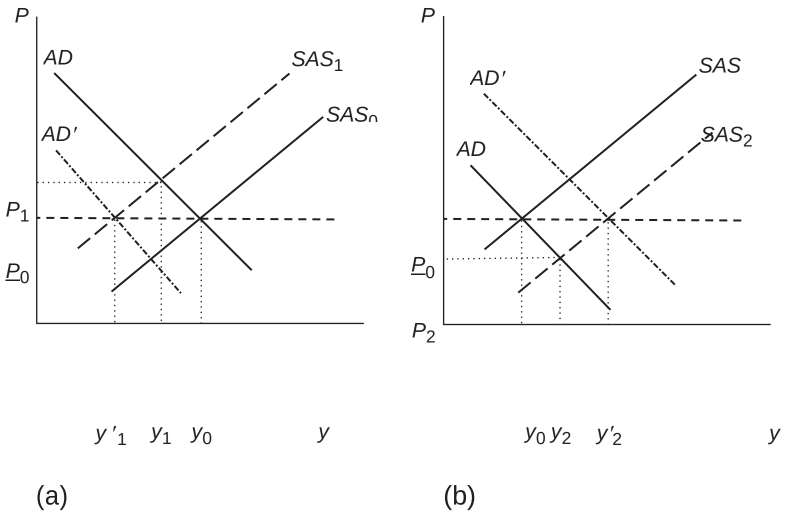


Figure 10.7

contractionary monetary policy would have increased the fall in output over that which would have occurred if the monetary policy had not been pursued.

Similarly, suppose that the aggregate supply shock had been a positive one, as shown in Figure 10.7b. This would shift the SAS curve to the right from SAS_0 to SAS_2, resulting in the increase in output from y_0 to y_2 and the decrease in prices from P_0 to P_2. The central bank could increase aggregate demand to stabilize the price level at P_0, but this would mean an expansionary monetary policy which shifts the AD curve to AD^J and further increases output to y_2^{J}. Price stabilization has, therefore, again increased the fluctuation in output.

Therefore, given the aggregate supply curve as being positively sloped and short run, the pursuit of price stability in the face of supply-side fluctuations has the cost of increasing the instability of output – and, therefore, of unemployment – in the economy. We leave it to the reader to adapt the analysis to the case of a vertical long-run supply curve.

Targeting the inflation rate

A low inflation rate, say in the 1 percent to 3 percent range, is generally considered to be effectively consistent with price-level stability, with the increase in prices merely reflecting the continual improvements in existing products and the introduction of new ones. Further, a positive but low rate of inflation is often considered to be beneficial for the economy, particularly in the labor market where it gives firms the flexibility to respond to shifts in the relative demand or supply of different products and types of workers, as well as shifts over time in the performance of a given worker. On the latter, firms can respond to small declines in productivity without having to reduce nominal wages, which creates industrial unrest, of workers whose real wage would fall. Inflation, as well as labor productivity increases, overcomes the societal norm of downward nominal wage rigidity. As against this beneficial so-called "grease effect" of inflation, errors in inflationary expectations can lead to a nominal wage being set in explicit and implicit labor contracts that result in a real wage higher or lower than the one that ensures full employment in the economy. This so-called "sand effect" occurs because of the two stages of wage negotiation and employment/production relevant to the derivation of the expectations-augmented Phillips curve (see Chapter 14). Such errors in inflationary expectations are less likely to occur with low, pre-announced and credible inflation targets than with high ones. Therefore, many central banks and economists generally believe that a low, pre-announced and credible inflation target improves the real performance of the economy in both the short and the long run.

Note that the inflation rate is not an operating target, since the monetary authority cannot directly change it. To maintain a target range for the inflation rate, the central bank will have to operate on the monetary aggregates and/or interest rates. Its success or failure will depend on the predictability of the relationships between the rate of inflation and these variables. Since the central banks of many countries have pursued a low inflation rate as a goal for more than a decade, a considerable amount of evidence has accumulated on it. This evidence shows that this goal has, in general, resulted in a reduction in the actual inflation rates. However, given the aggressive pursuit of this goal, this is not a surprising finding. However, as shown in the analysis above of price level targeting, targeting the price level alone tends to cause increased fluctuations in output and unemployment. This does not seem to have occurred in the past two decades, perhaps because central bank policies have followed not the single goal of price stability or a low inflation but a Taylor rule, which addresses both the output gap and the deviation of inflation from its target level. Chapters 11 and 15 also address this point.

A low inflation target versus a stable price-level target

Under the price-level target, if the actual price level falls below or rises above the target level, future policies would have to aim to bring it back to the target level. Hence, rising prices would have to be offset by future deflationary policies to make the price level return to its target level. Such a deflationary policy usually imposes costs in output and unemployment. By comparison, targeting the inflation rate allows the central bank to ignore one-time shifts in the price level, such as those due to changes in indirect tax rates, a shift in relative prices or an adjustment in the exchange rate, etc.

In addition, many economists believe that the public more easily relates to a low inflation rate target that remains constant over time and to the policies needed to maintain it, than to a price-level target and, in the presence of shocks, the inflationary and deflationary policies that may be needed to maintain the price target. This point becomes important since the transparency and credibility of policy is important to the public's expectations on inflation and the impact of monetary policy on the economy. Therefore, central banks have tended to adopt inflation targeting rather than price-level targeting. The popular Taylor rule embodies this preference for inflation rather than price-level targeting, with the target inflation rate that is usually set for developed economies being in the range from 1 percent to 3 percent.

Determination of the money supply

No matter how the money supply in the economy is defined or measured, several major participants are involved in its determination. They are:

1. The central bank, which, among its other policies, determines the monetary base and the reserve requirements for the commercial banks, and sets its discount rate.
2. The public, which determines its currency holdings relative to its demand deposits.
3. The commercial banks, which, for a given required reserve ratio, determine their actual demand for reserves as against their demand deposit liabilities.[11]

Some indication of the relative importance of the major contributors to changes in the money supply would be useful at this point. Phillip Cagan (1965) concluded that, in the USA, on average over the 18 cycles during 1877 to 1954, the fluctuations in the currency ratio had a relatively large amplitude over the business cycle. They caused about half of the fluctuations in the growth rate of the money stock, while fluctuations in the monetary base and the reserve ratio accounted for roughly one-quarter each. But, from a secular perspective, by far the major cause of the long-term growth of the money stock was the growth in the monetary base.

Therefore, there is considerable interaction between the behavior of the central bank, the public and the commercial banks in the money supply process. This interaction is important in studying the behavior of the central bank which, as a policy-making body deciding on the total amount of money desirable for the economy, must take into account the responses of the public and of the commercial banks to its own actions. The behavior of the central bank in the money-supply process becomes a distinctive topic of study, which is pursued later in this chapter and the next two chapters.

Demand for currency by the public

Fluctuations in the public's demand for currency relative to its holdings of demand deposits are a significant source of fluctuations in the money supply. The closest substitute – and a fairly close one at that – to currency holdings (C) is demand deposits (D), so that most studies on the issue examine the determinants of the ratio C/D, or of the ratio of currency to the total money stock (C/M1), rather than directly the determinants of the demand for currency alone.

The C/D ratio varies considerably, with a procyclical pattern over the business cycle and over the long term. The desired C/D ratio depends upon the individual's preferences in the light of the costs and benefits of holding currency relative to demand deposits. Some of these costs and benefits are non-monetary and some are monetary.

The non-monetary benefits and costs are related to the non-monetary costs of holding and carrying currency compared with those of holding demand deposits and carrying checks. They also take into account the general acceptability of coins and notes for making payments as against payments by other means. In financially less developed economies, with few bank branches in the rural areas and with banking usually not open to or economically feasible for lower income groups, even in the urban areas, currency has a clear advantage over checks. However, even in financially advanced economies, cash is almost always accepted for smaller amounts while the use of checks is restricted to payments where the issuer's credit-worthiness can be established, or the delivery of goods can be delayed until after the clearance of the check through the banks. It is also more convenient to make very small payments in cash than by writing a check. These aspects of non-monetary costs have changed substantially over time in favor of bank deposits with the expansion of the banking system and the modernization of its procedures, increasing urbanization, spread of banking machines, common usage of credit and debit cards, etc.

As against the greater convenience of currency over bank deposits for transactions, the possession of a significant amount of currency involves risks of theft and robbery, which impose not only its loss but also a risk of injury and trauma to the carrier. The fear of the latter is often sufficient to deter possession of large amounts of currency in societies where this kind of risk is significant. This is so in most countries, with the result that only small amounts of currency are carried by most individuals at one time or stored in their homes. By comparison, the demand for currency in Japan – a society with a very low rate of theft and robbery – is dominated by its convenience relative to bank deposits. Consequently, few individuals in Japan hold demand deposit accounts, checks are not widely accepted in exchange, even by firms, or given by them for payment of salaries. Many transactions, even of fairly large amounts, are done in currency.

The monetary costs and benefits of holding currency relative to demand deposits really relate to the net nominal return on the latter since currency does not have an explicit monetary return or service charge, while demand deposits often possess one or both of these. In any case, even if demand deposits pay interest, there is usually a negative return on them since banks incur labor and capital costs in servicing them and must recoup these through a net charge on them.[12]

Chapter 4 presented the inventory analysis of the demand for money. This model is applicable to the demand for currency relative to demand deposits. This was done in Chapter 4. As pointed out there, the problem with an application of this analysis that takes account only of the monetary cost of using currency versus demand deposits is that it ignores the non-monetary differences in their usage: acceptance in certain types of payments, risks of theft and robbery, etc. However, its central conclusion still holds: the optimal holdings of currency relative to demand deposits will depend on their relative costs and the amount of expenditures financed by them. Therefore, assuming both currency and demand deposits are "normal goods," an increase in the net cost of holding demand deposits would increase the demand for currency and hence the C/D ratio.

However, in a time-series context, the major reasons for changes in this ratio have been the innovations in shopping, payments and banking practices which have made checking increasingly easier and thereby lowered the C/D ratio. In addition, the significant possibility of theft and robbery – and the consequent risk to the person – with increases over time in many economies, have kept this ratio quite low or further reduced it. As indicated earlier, Japan, with a low risk from such criminal activities, is an exception to this rule and illustrates the greater convenience of using currency where a sufficiently wide range of denominations is made available in bank notes.

For the future, in financially developed economies, smart cards are likely to become a close substitute for currency in many transactions that used to be settled in currency since such cards may prove to be even more convenient than currency and yet not more susceptible to theft. Therefore, the demand for currency as a proportion of total expenditures or of M1 or M2 is likely to continue to decline in the future.

The above arguments imply that the demand function for currency can be written as:

$$\mathrm{C/D} = c(\gamma_{\mathrm{D}}, R^{\mathrm{h}}, R_{\mathrm{D}}, R_{\mathrm{T}}, Y; \text{payments technology}) \qquad (4)$$

where:

c = currency-demand deposit ratio
γ_{D} = charges on demand deposits
R^{h} = average yield on the public's investments in bonds, etc.
R_{D} = interest rate on demand deposits
R_{T} = interest rate on time deposits
Y = nominal national income

$\partial c/\partial \gamma_{\mathrm{D}} > 0$ and $\partial c/\partial R_{\mathrm{D}} < 0$ for obvious reasons. We expect $\partial c/\partial Y > 0$, since an increase in Y increases transactions that are likely to increase the demand for currency proportionately more than for demand deposits. This implies that the currency ratio will increase in the upturns and decrease in the recessions. An increase in the return on both time deposits and bonds is likely to decrease the demand for both currency and demand deposits. Further, currency is needed for small everyday transactions that tend to be inelastic in response to changes in interest rates, while efficient cash management techniques allow reductions in demand deposits. Hence, the currency ratio will rise with increases in R^{h} and R_{T}, so that $\partial c/\partial R^{\mathrm{h}} > 0$ and $\partial c/\partial T_{\mathrm{T}} > 0$. This implies that an increase in the rate of inflation and/or in the nominal interest rate, as usually happens in the upturn of the business cycle, would increase the currency ratio. Conversely, this ratio will fall in a recession. Hence, we expect the currency ratio to be procyclical (i.e. to rise in upturns and fall in downturns).

As stressed above, the currency ratio also depends upon the security environment and the availability of alternative modes of payment such as debit and credit cards. The impending innovations in creating smart cards that represent electronic purses are likely to reduce currency demand.

Households not only hold currency and demand deposits but also hold various forms of savings deposits, term deposits and their variants. All of these pay interest, and we can specify the arguments that would lead to the public's demand function for time deposits or for its desired ratio of time to demand deposits. The derivation of these functions is left to the reader.

10.6.2 Commercial banks: the demand for reserves

Commercial banks hold reserves against their deposits. A part of these reserves is normally held in cash (at the tills, in the automatic teller machines or in the bank's vault) and part is held in deposits with the central bank. If only a small part of deposits is withdrawn from a bank during a period, the bank does not have to maintain reserves equal to deposits (i.e. follow a 100 percent reserve ratio) but could increase its revenues by lending out a part or most of its deposits. This leads to *fractional reserve banking*, where the fraction of deposits held in reserves may be quite small, as discussed later in this chapter. The *reserve ratio* is the ratio of reserves held to deposits.

The central bank often requires the commercial banks to meet a certain minimum ratio called the *required reserve ratio*[13] – of their reserves to their deposit liabilities.[14] Chapter 11 will present the required reserve ratios for several countries. In 1999, this ratio was zero in Canada and the United Kingdom. In the United States, it depended on the amount of deposits and varied between 3 percent and 9 percent for depository institutions.

Banks usually hold reserves in excess of those required to meet the required reserve ratio. Banks also borrow from other banks or the central bank. Reserves held in excess of the sum of required and borrowed reserves are referred to as *free reserves* – that is, at the bank's disposal for use if it so desires.

Free reserve hypothesis

Free reserves for a bank are those it wants to hold in addition to its required reserves and borrowed reserves. Free reserves must be distinguished from *excess reserves*, which are actual holdings of cash reserves in excess of the sum of required, borrowed and free reserves. Excess reserves are ones that the bank wants to eliminate either immediately or gradually. The hypothesis for the determination of free reserves is known as the *free reserve hypothesis.*

Required reserves and free reserves depend upon the required reserve ratio or differential ratios imposed by the central bank, discussed in Chapters 11 and 12 on central bank behavior, and upon the total deposits in the bank. The computation of such required reserves is largely mechanical, according to a formula prescribed by the central bank.[15]

Each bank has to anticipate its deposit liabilities in making its decisions on its reserve holdings. Demand deposits may be withdrawn on demand at any time and individuals' demand deposits in any given bank fluctuate considerably over time as they make deposits and withdrawals. For any given bank over a given period, the totals of new deposits and withdrawals are likely to cancel out to some extent, depending upon its size and the distribution of its depositors among occupations and industries, between employees and employers, etc. The degree of uncertainty as to the average levels of deposits in any bank is thus likely to vary between banks. It is likely to be higher for unit rather than branch banks, small rather than large banks, and banks with a smaller degree of monopoly than those with a greater one. The cancellation process is likely to be still greater for all the banks in the economy taken together, so that the overall amounts of demand deposits in the economy normally exhibit a great deal of stability.

For banks, reserves are an asset, in addition to bonds and loans, so that the demand for reserves depends on the returns on the latter. Reserves do not generally earn a monetary return. Their demand should, because of the substitution effect, fall as the rates of return on the other assets rise, and vice versa.

Under uncertainty, the free reserve hypothesis assumes that banks maximize the expected utility of their terminal wealth, corresponding to the expected utility maximization by the individual investor presented in Chapter 5. Therefore, the theory of portfolio selection set out in Chapter 5 can be adapted to explain the bank's demand for free reserves. Assuming that the banks are risk averters, disliking the prospect of ending up with less than the required reserves, they would always hold more than the required reserves. Part of these extra reserves could be borrowed, so that the demand for free reserves will depend upon the risks present, the response to risk, the cost of borrowing and the return on other assets in the bank's portfolio.

A significant element of the risk in falling short of the desired reserves depends on the structure of the banking system, the size of the bank in question and the diversity of its client base. Canadian and British banks tend to be large, with branches all over the country. They have a very diversified client base, so that the daily variance in their deposits is relatively low. The US banks in the smaller cities and rural areas are often small, have a limited number of branches and may be dependent on a particular segment of the economy. Consequently, they face higher daily variance in their deposits.

Another significant element of risk in falling short of the desired reserves is related to the formula specified by the central bank for the minimum reserves that the banks should hold against their deposits.[16] In the UK, although the reserve requirement is zero (or, rather, non-negative), the banks are expected to meet it on a daily basis. This increases their risk, which to some extent is offset through their ability to borrow reserves in the overnight market from other financial institutions. Canada allows averaging of reserves and deposits over a 4 to 5-week period and the United States allows this over 2 weeks, thereby implying less risk for their banks than for British banks.[17]

Borrowing by commercial banks from the central bank

Banks borrow reserves from a variety of sources. Banks frequently borrow reserves from each other bilaterally and often do so in the context of an organized overnight loan market such as the Federal Funds market in the USA and the Overnight Loan market in Canada. They can, depending on regulations, also borrow abroad to supplement their reserves. Borrowing by individual banks from other banks within the system does not change the monetary base and can therefore be ignored in the determination of the money supply. However, when the commercial banks as a whole increase their borrowing from the central bank or from abroad, the monetary base increases and the money supply expands.

In lending to the commercial banks as a whole, the central bank is said to act as the lender of last resort since the banking system as a whole cannot obtain additional funds from its own internal borrowing and lending. However, individual commercial banks sometimes treat the central bank as their lender of first resort rather than last resort. The terms on which it lends and the conditions under which its loans are made affect the amounts that the banks wish to borrow from it. In general, borrowing from the central bank can trigger greater oversight by the central bank into the borrowing bank's asset management and other practices. Since this is rarely desired, it acts as a disincentive to borrow.

In the USA, the discount rate – at which the Federal Reserve System lends to its member banks – is usually below the three-month Treasury bill rate, so that these banks stand to gain by borrowing from the Federal Reserve System. To limit the amounts and frequency of borrowing, the Federal Reserve Board imposes a variety of formal and informal rules on such borrowing. One of the latter is that borrowing from the Federal Reserve System is a *privilege*, extended by the Federal Reserve System to its member banks, rather than a right of the banks. This privilege can be curtailed or circumscribed by conditions if a bank tries to use it indiscriminately.

Canada has experimented with two different methods for setting the bank rate at which it lends to the chartered banks. Under a fixed bank rate regime adopted from 1956 to 1962 and from 1980 to 1994, the bank rate was automatically set each week at 0.25 percent higher than the average Treasury bill rate that week. Since it was higher than the Treasury bill rate, the chartered banks incurred a loss if they financed their purchases of Treasury bills by borrowing from the Bank of Canada. Such a rate is said to be a "penalty rate" and, by its nature, discourages borrowings. It does not therefore need the support of other restrictions on borrowings to the extent that the American discount rate requires. Since 1994, the Bank has placed its major focus on setting the overnight loan rate, with an operating band around it of 50 basis points. This rate is the rate at which the banks and other major participants in the money market make overnight loans to each other. Since 1996, the bank rate has been set at the upper limit of the operating range for the overnight loans, so that it is a floating rate and continues to be a penalty rate, irrespective of daily movements in the market rates. The Bank of Canada influences the Bank Rate by changes in its supply of funds to the overnight market or through its purchases or sales of Treasury bills.

In the United Kingdom, the Bank of England determines daily the rate at which it will lend to the banks. This allows it control over borrowings from it on a daily basis. It also allows close control over the interest rates that the banks charge their customers, since these rates have base rates closely tied to the Bank's daily rate.

Banks' demand function for reserves

The free reserve hypothesis, in attempting to explain the demand for free reserves, has to take these differing practices into account and would yield differing demand functions for different countries and periods. However, empirical studies[18] have confirmed its implication that the earnings on alternative assets influence the amount of reserves held and that the ratio of desired reserves to demand deposits cannot be taken as constant for purposes of monetary policy.

The preceding arguments imply that the demand function for desired free reserves, FR, can be expressed as:

$$\mathrm{FR}/D = f\,(R, R_{\mathrm{BR}}, R_{\mathrm{CB}}) \tag{5}$$

where:

R = average interest rate on banks' assets
R_{BR} = average return on banks' reserves
R_{CB} = central bank's discount rate (for loans to the commercial banks)

In (5), $\partial f/\partial R < 0$, $\partial f/\partial R_{\mathrm{BR}} > 0$ and $\partial f/\partial R_{\mathrm{CB}} > 0$. We have simplified this function by including average interest rates rather than the variety of interest rates that will need to be considered in practice.

Note that, in (5), R is an average of the nominal returns on Treasury bills, bonds of different maturities, mortgages and loans to the public, etc. Also note that R_{BR} would be zero for reserves held in currency and would also be zero if the rest of the reserves were held in non-interest paying deposits with the central bank. Banks will want to increase free reserves if the cost R_{CB} of covering a shortfall in reserves increases and decrease them if the return R from investing their funds rises. Free reserves would also increase if R_{BR} were positive and were to increase.

Mechanical theories of the money supply: money supply identities

Mechanical theories of the money supply are so called because they use identities, rather than behavioral functions, to calculate the money supply. The money-supply equations resulting from such an approach can be easily made more or less complex, depending upon the purpose of the analysis. We specify below several such equations, starting with the most elementary one.

An elementary demand deposit equation

Assume that the ratio of reserves held by the banks against demand deposits is given by:

$$BR = \rho D$$

where:

BR = reserves held by banks
D = demand deposits in banks
ρ = reserve ratio

If ρ equals the required reserve ratio, set by the central bank for the banking system, and BR represents the reserves exogenously supplied to it, profit-maximizing banks will create the amount of deposits given by:

$$D = (1/\rho)BR \tag{6}$$

This equation is the elementary deposit creation formula for the creation of deposits by banks on the basis of the reserves held by them. It suffers from a failure to take note of the behavior of the banks and the public in the deposit expansion process, so that a more elaborate approach to the money supply is preferable.

Common money-supply formulae

Friedman and Schwartz (1963) used a money-supply equation that not only takes account of the reserve/deposit ratio of banks but also of the ratio of currency to deposits desired by the public. Their ratio is derived simply from the accounting identities:

$$M = C + D \tag{7}$$

$$\mathrm{M0} = BR + C \tag{8}$$

where:

D = demand deposits
BR = banks' reserves
C = currency in the hands of the public
M0 = monetary base = $BR + C$.

The Friedman and Schwartz money-supply formula is derived as follows:

$$\begin{aligned} M &= \frac{M}{\mathrm{M0}}\mathrm{M0} \\ &= \frac{C + D}{BR + C}\mathrm{M0} \\ &= \frac{1 + D/C}{BR/C + 1}\mathrm{M0} \\ &= \frac{(1 + D/C)(D/BR)}{(BR/C + 1)(D/BR)}\mathrm{M0} \\ &= \frac{(1 + D/C)(D/BR)}{(D/BR + D/C)}\mathrm{M0} \end{aligned} \tag{9}$$

Equation (9) separates the basic determinants of the money stock into changes in the monetary base and changes in the "*monetary base multiplier*," defined as ($\partial M/\partial$M0), for the monetary base.[19] This multiplier is itself determined by D/BR, the reserve ratio, and C/D, the currency ratio. Of these, the reserve ratio reflects the required reserve ratio and the banks' demand for free reserves. The currency ratio reflects the public's behavior in its demand for currency. Hence, the three determinants of the money stock emphasized by (9) are the monetary base, the currency and the reserve ratios.

Another version of the money-supply formula is:

$$M = \left[\frac{1}{\frac{C}{M} + \frac{BR}{D} - \frac{C}{M}\cdot\frac{BR}{D}}\right]\mathrm{M0} \tag{10}$$

Using this equation, Cagan (1965) examined the contributions of the three elements B, C/M and BR/D, to M2 over the business cycle and in the long term. He found that the dominant factor influencing the long-term growth in the money stock was the growth in the monetary base. Changes in the two ratios contributed little to the *secular* change in M2. However, for *cyclical* movements in the money stock the changes in the C/M ratio were the most important element, whereas the reserve ratio had only a minor impact and changes in the monetary base exerted only an irregular influence.

The currency–demand deposit – and hence the currency–money ratio – is influenced strongly by changes in economic activity and especially by changes in the rate of consumer spending. As we have explained in earlier sections, this ratio varies in the same direction as nominal national income – hence, pro-cyclically – so that a rise in spending in cyclical upturns increases currency holdings, which lowers the money supply.

The preceding money-supply formula does not differentiate deposits into various types such as demand deposits, time and savings deposits, and government deposits nor does it differentiate between their reserve requirements. A formula (its derivation is not shown) that does so is:

$$M^{\Sigma} \quad \frac{1+c \qquad \Sigma}{\rho_D + \rho_T t + \rho_G g + c} \underset{=}{} \mathrm{M0} \qquad (11)$$

where t T/D and g G/D, T and G represent time/savings deposits and government deposits respectively in commercial banks.

The above formulae are all identities. Which one is used depends upon the rules and regulations about reserve ratios, the availability of statistical data and the further behavioral assumptions that are made. In practice, theories of the money supply go beyond these identities and embed the relevant identity in a behavioral theory.

Behavioral theories of the money supply

A behavioral theory of the money supply process must take into account the behavior of the different participants in this process in order to determine the economic and non-economic determinants of the variables being studied. Such a theory studies this behavior in terms of the main components of the preceding money supply formulae, such as the currency desired by the public, the reserves desired by the commercial banks, the amounts borrowed by them and the monetary base which the central bank wishes to provide.

Our earlier discussions on the demand for components of the money supply imply the general form of the money-supply function to be:

$$M^s = M^s(R_D, R_T, R_S, R_L, R_d, R_O, R, Y, \mathrm{M0}) \qquad (12)$$

where:

R_D = charges on demand deposits
R_T = interest rate on time deposits
R_S = short-term interest rate
R_L = long-term interest rate
R_d = discount rate (central bank rate for lending to the commercial banks)
R_O = overnight loan rate
RR = required reserve ratio.

As argued in earlier sections, the money supply depends, among other variables, upon the free reserves desired by the banks and the currency desired by the public. Free reserves depend upon R_O, R_d, R_S and R_L since these determine the opportunity cost of holding free reserves. The public's demand for currency will similarly be a function of R_D and R_T. It also depends, as argued earlier, on the level of economic activity for which nominal national income Y is a proxy. Finally, the money supply depends upon the monetary base M0.

The monetary base is under the control of the monetary authorities, which can operate it in such a way as to offset the effect of changes in the other variables on the money supply. Alternatively, they may only allow the changes in the other variables to affect the money supply in so far as the effect changes the money supply to a desired extent. The monetary base is, therefore, not necessarily a variable independent of the other explanatory variables in (12).

Now consider the directions of the effects that are likely to occur in the money-supply function. An increase in the monetary base increases the money supply. An increase in national income increases the currency demand and lowers banks' reserves and hence decreases the money supply. An increase in the short-term market interest rate increases the profitability of assets which are close substitutes for free reserves and hence decreases the demand for free reserves, which increases the money supply. A cut in the discount rate has a somewhat similar effect. An increase in the yield on time deposits increases their demand by the public, which lowers the reserves available for demand deposits, so that demand deposits decrease.

In practice, the estimation of the money-supply function is usually undertaken with a smaller number of variables than those specified in (12). This is partly because of collinearity among the various interest rates.

Equation (12) specifies the money-supply function and could be applied to the supply of either M1 or M2 or another monetary aggregate. However, note that the signs of the interest-rate elasticities could differ among the aggregates. There are three main cases of differences:

1 If the interest rates on demand deposits increase, their demand will increase but this could be merely at the expense of time deposits, so that while M1 increases, M2 does not do so. Hence, the interest elasticity of demand deposits and M1 with respect to R_D is positive but that of M2 may be positive or zero.

2 Since time deposits are excluded from M1, they are part of the opportunity cost of holding M1. Hence, the elasticity of M1 with respect to the interest rate on time deposits, R_T , would be negative. These deposits are, however, part of M2 so that, when their interest rate increases, the desired amount of time deposits increases and so does M2. That is, the interest elasticity of M2 with respect to R_T is positive.

3 The interest elasticity of M1 would be negative with respect to the return on bonds. But if the time deposits interest rates rise with this bond rate, the respective interest elasticity of M2 is likely to be zero. However, if the time deposit rates do not increase when the bond rate rises, the public will switch some time deposits to bonds, so that the elasticity of M2 with respect to R would be negative. Therefore, this interest elasticity depends upon the relationship between R_T and$R$.

Table 10.2 shows the pattern of interest-rate elasticities for M1 and M2. This table also includes the elasticities of M1 and M2 with respect to the return on excess reserves and the central bank discount rate on borrowed reserves. Both are negative. The reasons for these and the other signs shown in Table 10.2 have been explained above.

Interest rate elasticities of the money supply

Empirical studies on money supply are far fewer than on money demand. The following brief review of the empirical findings on the money supply function confines itself to reporting the elasticities' estimates for this function.

Rasche (1972) reports the impact and equilibrium elasticities calculated by him for the supply functions reported by DeLeeuw (1965) for the Brookings model, by Goldfeld (1966) for the Goldfeld model, and for the MPS model developed by the Federal Reserve–MIT–Pennsylvania econometric model project. These studies were for the USA, using data up to the mid-1960s. Rasche's calculations of these elasticities are roughly in the ranges reported in Table 10.3.[20]

Note that there have been many changes in the United States' financial markets since the 1960s, so that the elasticity ranges reported in Table 10.3 are now mainly useful for pedagogical purposes. These elasticities indicated that the main components of the money supply – and the money supply itself – were not exogenous but functions of the interest rates in the economy. We conclude from Table 10.3 that:

1 Currency demand was negatively related to the time deposit rate, whereas time deposits were positively related.

2 Time deposit holdings were positively related to their own interest rate and negatively to the Treasury bill rate.

3 Banks increased their borrowing from the Federal Reserve as the Treasury bill rate rose and decreased them as the discount rate increased. Note that the discount rate is the cost of such borrowing and the Treasury bill rate represents the return on funds invested by the banks. Hence an increase in the Treasury bill rate provides an incentive for banks to increase their loans for a given monetary base, and also to increase their borrowing, given the discount rate, from the central bank. Both these factors imply a positive elasticity of the money supply with respect to the Treasury bill rate, as shown in Table 10.3.

4 Conversely, while the banks' free reserves decreased as the Treasury bill increased, they increased with the discount rate. The Treasury bill rate represents the amount the banks lose by holding free reserves and is, therefore, their opportunity cost, so that the free reserves fall with the Treasury bill rate. However, these elasticities with respect to the discount rate are positive since the discount rate is the "return" on free reserves; i.e. if they hold adequate free reserves to meet withdrawals, the banks escape having to borrow from the central bank at the discount rate.

These reported elasticities are consistent with the analysis presented earlier in this chapter. Even though there have been numerous innovations in the financial markets since the 1960s, so that the magnitudes of the actual elasticities are likely to have changed, there is no reason to expect that the *signs* of elasticities have altered from those reported above.

Lags in the money-supply function

The main findings on this show that:

1 The impact elasticities are significantly lower than the equilibrium ones, indicating that the adjustments take longer than one quarter.
2 The money supply had positive interest elasticities each month through the first 18 months for which the elasticities were reported.

These findings show that the financial sector did not adjust the money supply to its full equilibrium level within one quarter. In fact, the second finding points out that the money supply continues to change even after six quarters. This finding has been confirmed by many studies, so that the existence of lags in the response of the money supply to interest-rate changes can be taken as being well established.

Cointegration and error-correction models of the money supply

There are few cointegration studies on the money supply function and its major components. We draw the following findings from Baghestani and Mott (1997) to illustrate the nature of empirical findings on money supply and the problems with estimating this function when monetary policy shifts.

Baghestani and Mott performed cointegration tests on USA monthly data for three periods, 1971:04 to 1979:09, 1979:10 to 1982:09 and 1983:01 to 1990:06, using the Engle–Granger techniques. Their variables were log of M1, log of the monetary base (*B*) and an interest rate variable (*R*). The last was measured by the three-month commercial paper rate for the first two periods and by the differential between this rate and the deposit rate paid on Super NOWs (Negotiable Orders of Withdrawal at banks) introduced in January 1983. Further, the discount rate was used as a deterministic trend variable, since it is constant over long periods. The data for the three periods was separated since the Federal Reserve changed its operating procedures between these periods.

Baghestani and Mott could not reject the null hypothesis of no cointegration among the designated variables for 1971:04 to 1979:09. Further, for 1979:10 to 1982:09, while M0 and *R* possessed a unit root, M1 did not, so the cointegration technique was not applied for this period. The only period which satisfied the requirement for cointegration and yielded a cointegration vector was 1983:01 to 1990:06. The error-correction model was also estimated for this period. The cointegration between the variables broke down when the period was extended beyond 1990:06. These results have to be treated with great caution. As indicated in Chapter 9 on money-demand estimation, cointegration is meant to reveal the long-run relationships and, for reliable results, requires data over a long period rather than more frequent observations, as in monthly data, over a few years. The three periods used by Baghestani and Mott were each less than a decade.

For 1983:01 to 1990:06, Baghestani and Mott concluded from their cointegration–ECM results that the economy's adjustments to the long-run relationship occurred through changes in the money supply and the interest rate, rather than in the monetary base. Comparing their findings across their three periods, we see that changes in the central bank policy regime, such as targeting monetary aggregates or interest rates, are extremely important in determining the money supply function in terms of both its coefficients and whether there even exists a long-run relationship. Further, even regulatory changes such as permitting, after 1980, the payment of interest on checkable deposits can shift the money-supply function.

Monetary base and interest rates as alternative policy instruments

The central bank may use either the monetary base or the interest rate as a way of controlling aggregate demand in the economy, or may have to use both. Under certainty and known money supply and demand functions, it needs to use only one of them since the use of either of them indirectly amounts to use of the other. To see this correspondence, assume that the money supply function is given by:

$$M = \Sigma \frac{\mathrm{M0}}{\frac{C}{M} + \frac{BR}{D} - \frac{C}{M} \cdot \frac{BR}{D}} \Sigma \qquad (13)$$

where the meanings of the symbols are as explained earlier. This money supply function can be written simply as:

$$M = \alpha \mathrm{M0}, \tag{14}$$

where:

$$\alpha = \Sigma \frac{1}{\frac{C}{M} + \frac{BR}{D} - \frac{C}{M} \cdot \frac{BR}{D}} \Sigma$$

where α is the "monetary base (to money supply) multiplier" $\partial M/\partial \mathrm{M0}$, though some authors call it the "money multiplier," a term that we have more appropriately reserved for $\partial Y/\partial M$, where Y is nominal national income.

Let the general form of the money demand function be:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R) \tag{15}$$

where R is the nominal interest rate and y can be the pre-specified actual or desired level of output. In money market equilibrium, we have:

$$\alpha.\mathrm{M0} = P.m^{\mathrm{d}}(y, R) \tag{16}$$

Under certainty, given the policy targets for P at P^* and y at y^*, (16) can be solved for the relationship between M0 and R, so that the central bank can achieve its objectives by setting the monetary base M0 at $\mathrm{M0}^*$ and letting the economy determine R, or by setting R at R^* and letting the economy determine the money supply needed to support R^*. It does not have to pursue a policy of setting both M0 and R.

Making the choice between the interest rate and the monetary base as operating targets in a stochastic context

In developed economies, adequate reasons for the choice between M0 and R as the optimal monetary policy instrument arise only if there is uncertainty and unpredictability of the money supply and/or demand functions.[21] In this scenario, the policy maker may not know which policy instrument would more predictably deliver the target values y^* and P^*. In general, in the absence of any sure information, the theory of policy under uncertainty implies that the risk-averting policy maker should diversify by using both policy instruments. However, as Poole's analysis presented earlier in this chapter (and in Chapter 13) shows, if the main shocks to aggregate demand originate from shifts in the commodity sector, then the money supply is, in general, the preferable monetary policy instrument. But, if the main shocks to aggregate demand originate in the monetary sector, the interest rate is, in general, the preferable one (Poole, 1970). The next chapter examines this issue in greater detail.

Another reason for the choice between the two instruments can arise because of the nature of the economy. Underdeveloped financial economies usually do not have well-developed bond markets, which prevents the central bank from effectively using open-market operations. However, there may be other ways of changing the monetary base, such as by feeding increases in the money supply to the government to finance its deficits. In addition, changes in the reserve requirements can be used to change the supply for a given monetary base. Such economies also have fragmented financial markets, along with a large informal financial sector, so that there need not be a close relationship between the interest rate set by the central bank and that charged in the various private financial markets. Further, much of the investment in these economies may not be sensitive to market interest rates. Therefore, on the whole for such economies, it is likely that changes in the money supply will be more effective, though not perfectly so, in manipulating aggregate demand than will changes in the interest rate. However, given the imperfections of both policy instruments in controlling aggregate demand, demand management would work better if the central bank were to use both instruments.

Chapter 1 presented a brief definition of the classical paradigm and its component models. Chapter 2 covered the heritage of monetary and macroeconomic theory, and in Chapter 3 we presented microeconomic analyses of the demand and supply functions for commodities, money and labor. These chapters should be reviewed at this stage.

Chapter 13, on the determination of aggregate demand, is a prerequisite for this chapter, which takes the determination of aggregate demand as a given for its analysis.

As was pointed out in Chapter 1, there are two major paradigms in short-run macro modeling: classical and Keynesian. The classical paradigm encompasses the traditional classical set of ideas, the neoclassical model, the modern classical model and the new classical model. This chapter starts with the neoclassical model and returns to the other two models towards the end.

As in earlier chapters, lower case symbols will generally refer to the real values of the variables and upper case symbols to their nominal values.

Stylized facts on money, prices and output

Monetary macroeconomics is vitally concerned with the empirical relationships between monetary policy, prices, inflation and output. For a macroeconomic theory to be valid and useful, it must do a reasonable job of explaining these relationships. The stylized facts on these relationships are:

1 Over long periods of time, there is a roughly one-to-one relationship between the money supply and inflation (McCandless and Weber, 1995).
2 Over long periods, the relationship between inflation and output growth is not significant or, if significant, is not robust (Taylor, 1996).[1] This is also so for money growth and output growth (Lucas, 1996; Kormendi and Meguire, 1984; Geweke, 1986; Wong, 2000), though some studies show a positive correlation between these variables, especially for low-inflation countries (McCandless and Weber, 1995; Bullard and Keating, 1995), while others show a negative relationship (Barro, 1996).
3 Over long periods of time, the correlation between money growth rates and nominal interest rates is very high (about 0.7 or higher) (Mishkin, 1992; Monnet and Weber, 2001), so that changes in interest rates tend to reflect changes in inflation.
4 In the short run, changes in money supply and interest rates have a strong impact on aggregate demand (Anderson and Jordan, 1968; Sims, 1972).
5 Changes in money growth lead to changes in real output in business cycles (Friedman and Schwartz, 1963a, b). Unanticipated money supply changes affect output (Barro, 1977), as do anticipated ones (Mishkin, 1982) (see Chapter 9). Negative shocks to money supply have a stronger impact on output than positive ones (Cover, 1992).
6 Monetary policy increases (decreases) in the short-term interest rate lead to a decline (an increase) in output (Eichenbaum, 1992).
7 On monetary policy dynamics in the short run, monetary shocks build to a peak impact on output and then gradually die out, so that there is a “hump-shaped pattern” of the effect of monetary policy on output (see Mosser (1992), Nelson (1998) and Christiano *et al*. (1999) for evidence on the United States, and Sims (1992) for evidence for some

other countries) with the peak effect occurring with a lag longer than one year, sometimes two to three years.

8 As a corollary of the preceding point, the impact of monetary shocks on prices occurs with a longer lag than on output, so that the impact of monetary shocks on output does not mainly occur through prior price movements.

9 For the short run, since inflation responds more gradually than output to monetary policy changes, expected inflation also responds more gradually. Therefore, errors in price or inflation expectations do not provide a satisfactory explanation of the response of output to monetary policy shifts.

10 Unanticipated price movements explain only a small part of output variability. The impact of unanticipated, as well as anticipated, monetary shifts on real output do not necessarily occur through prior price/inflation increases (Lucas, 1996, p. 679).

11 The responses of output and prices to monetary shocks differ over different episodes. They are also stronger for contractionary than for expansionary monetary episodes.

12 Contractionary monetary policies to reduce inflation do initially reduce output significantly (Ball, 1993), often for more than a year. The cost in terms of output tends to be larger if inflation is brought down gradually rather than rapidly. It is lower if the policy has greater credibility (Brayton and Tinsley, 1996).

Definitions of the short run and the long run

Equilibrium in a model is defined as that state in which all markets clear, so that the demand and supply in each market are equal. Another definition is that it is the state from which there is no inherent tendency to change. The classical paradigm uses the former definition. The latter definition is met when the former one is satisfied.

The *long run* of the short-run macroeconomic model is defined as that *analytical* period in which:

1 There are no adjustment costs, inertia, contracts or rigidities, so that all adjustments to the desired or equilibrium values of its variables are instantaneous.

2 There are no errors in expectations, so that the expected values of the variables are identical with their actual values. This condition is trivially satisfied if there is certainty.

3 There exists long-run equilibrium in all markets.

Note that these assumptions imply that there are no labor contracts between firms and workers and no price contracts among firms, so that prices and nominal and real wages adjust instantly and fully to reflect market forces. However, the long run of the short-run macroeconomic model still assumes that the capital stock, labor force and technology are constant.

Given these assumptions, the economy's resulting long-run employment level is said to be the "full-employment" level and its long-run output is said to be "full-employment output," for which our symbol is y^f. Note that the long-run – or full-employment – output is not really the maximum output that the economy could produce at any time, e.g. if all its resources were used 24 hours a day. It is also not the equilibrium level of output that would be produced in the short run or the actual output that might be produced when the economy is not in equilibrium. Hence, macroeconomics interprets the term "*fully employed*" in a special way. It is formally defined as being the level of output and employment that would exist in long-run equilibrium of the short-run macroeconomic model. Intuitively, this corresponds to the levels of output and employment that can be sustained by the economy over the long run with its current supplies of the factors of production and its current technology – given its current economic, political and social structures, as well as the wishes of the owners of the factors of production.

By comparison with the definition of the long run, the *short run* in the context of the short-run macroeconomic model is defined as that analytical period in which:

1 Some variables, especially the capital stock, technology and labor force, are constant.
2 There can be adjustment costs, e.g. of adjusting prices, wages, employment and output to their desired levels, as well as inertia, contracts or/and rigidities of some kind.
3 There can be errors in expectations; e.g. the expected values of variables such as prices, inflation, wages and aggregate demand, differ from their actual values.
4 There can be disequilibrium in any or all of the markets.
5 The short-run equilibrium values of the various variables can differ from their long-run values. In particular, the short-run output can be greater or less than its full-employment level.

The *actual* values of the variables in the economy can differ from their long run and short-run equilibrium levels for a variety of reasons. This would occur if the actual economy is not even in short-run equilibrium or suffers short-run deviations from full employment due to factors other than errors in price expectations. Note that actual output occurs in a chronological time period whereas the short run and the long run are analytical, hypothetical, constructs. Illustrating this point by reference to the output of commodities, there are three concepts of output: actual output, short-run equilibrium output and long-run equilibrium (full-employment) output.

Also, note that the terms “short run” and “long run” are different in meaning from their counterparts of “short period” or “short term” and “long period” or “long term.” The former are analytical constructs indicative of the forces allowed to work in the analysis; the latter are chronological ones and refer to an interval of actual time. In the real world, the economic forces encompassed in both the analytical short run and the long run operate simultaneously at every moment of time in the economy. To illustrate, the population and the capital stock are continuously changing, so that the analytical forces of the long-run growth models must be continuously operating in the economy, even over the next day, month or quarter.[2] At the same time, the analytical forces of the short-run macroeconomic models are also simultaneously operating in the economy.

Long-run supply side of the neoclassical model

Production function

In industrialized, as against agricultural, economies, capital and labor are the dominant inputs in production, while land plays only a minor role and is normally not included in macroeconomic analysis. With this assumption, the production of commodities is taken to use only labor and capital as inputs. The production function for the economy – as represented by that for the representative firm – can then be written as:

$$y = y(n, K) \quad y_n, y_K > 0; y_{nn}, y_{KK} < 0 \tag{1}$$

where:

y = output
K = physical capital stock
n = labor employed.

The physical stock of capital has already been assumed to be constant in the short-run macroeconomic context of our analysis, so that $K = \underline{K}$, where the underlining indicates "constancy" or "exogeneity," as required by the definition of the short run. Hence, we have:

$$y = y(n, \underline{K}) \quad y_n > 0, y_{nn} < 0 \tag{2}$$

With this modification, labor is left as the only variable input, with a positive relationship between output and employment. The assumption of $y_n > 0$ and $y_{nn} < 0$ is that the marginal productivity of labor is positive but diminishing: successive increments of labor yield smaller and smaller increments of output.

Labor market in the long run

The specification of the labor market requires specification of the demand and supply functions of labor and its equilibrium condition. Their derivation was presented in Chapter 3. We here present the simplified derivations used in standard neoclassical macroeconomic models, which imply that the demand and supply of labor depend only on the real wage rate. However, intertemporal analysis implies that both these functions will also depend on the real interest rate and future wage rates. On this issue, empirical studies of both labor demand and supply analysis show that, for short periods, neither significantly depends on the interest rate or the future wage rates for the ranges in which these variables normally fluctuate. Therefore, allowing empirical findings to determine the relevant theoretical assumptions, the following macroeconomic model has labor demand and supply functions that depend only on the (current) real wage.

Production analysis assumes that firms maximize profits and operate in perfectly competitive markets. Hence, they employ labor until its marginal revenue product equals its nominal wage rate. Using the price level to divide both of these variables, in perfect competition and for the representative firm, profit maximization requires that firms employ labor up to the point where the real value of its marginal product equals its real wage rate. That is,

$$y_n(n, \underline{K}) = w \tag{3}$$

where:

y_n = marginal product of labor
w = real wage rate.

Since K is held constant at $\underline{K}$, it is omitted from further analysis. Solving (3) for employment n, and designating this value as the demand for labor n^d by firms:

$$n^d = n^d(w) \quad \partial n^d/\partial w < 0 \tag{4}$$

Chapter 3 derived the supply function of labor from utility maximization subject to a budget constraint. Its simple version in a single commodity world is:

$$n^{s} = n^{s}(w) \quad \partial n^{s}/\partial w > 0 \tag{5}$$

where n^{s} is the supply of labor. Note that (5) specifies that the supply of labor depends upon the real rather than the nominal wage. Hence, workers are free from *price illusion*, which is the distortion caused when money wage rates and the prices of commodities rise by identical proportions but the workers, looking at the rising nominal wage rates, believe that they are better off even though the purchasing power of wages has remained unchanged.

Long-run equilibrium levels of employment and output

It is important to note that there are no adjustment costs or errors in expectations in the preceding framework, so that its equilibrium will be the long-run equilibrium, as against a short-run equilibrium in a model that allows for adjustment costs and expectational errors.

Equilibrium in the labor market requires that:

$$n^{d}(w) = n^{s}(w) \tag{6}$$

Since (6) is an equation in only one variable, w, solving it would yield the long-run equilibrium wage rate w^{LR}. This wage rate, substituted in either the demand or the supply function, yields the long-run equilibrium level of employment n^{LR}. This level of employment, substituted in the production function (2), yields the long-run equilibrium level of output y^{LR} for the economy.

Note that n^{LR} equals both the demand and supply of labor at the equilibrium wage. Therefore, at n^{LR}, all the workers who want jobs at the existing wage rate are employed and the firms can get all the workers that they want to employ. This is the definition of full employment,[3] so that the equilibrium level of employment n^{LR} represents full employment and, to emphasize this property, can be designated as n^{f}. Its corresponding long-run equilibrium output level y^{LR} is the full-employment level of outputy^{f}.

These conclusions on the labor market equilibrium are that:

$$n = n^{LR} = n^{f} \tag{7}$$

From (2) and (7), we have:

$$y = y^{LR} = y^{f} \tag{8}$$

Diagrammatic analysis of output and employment in the neoclassical model

Figure 14.1 plots the demand and supply functions of labor, with the usual slopes for demand and supply curves. Equilibrium occurs at (n^{LR}, w^{LR}). Figure 14.2 plots the

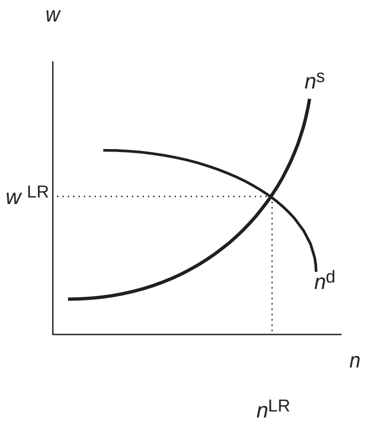


Figure 14.1

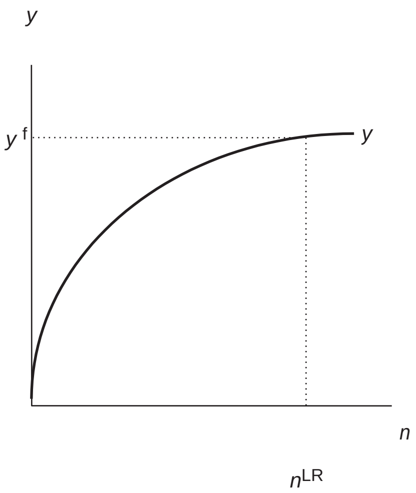


Figure 14.2

production function: output y is plotted against employment n and the curve is labeled y (output). This curve has a positive slope and is concave, representing the assumption of the diminishing marginal productivity of labor. The equilibrium level of employment n^{LR} determined in Figure 14.1 is carried over into Figure 14.2 and implies the equilibrium – and profit-maximizing – output y^{f}.

From (7) to (8) – and as shown in Figures 14.1 and 14.2 – the equilibrium levels of the real wage rate, employment and output do not depend upon the price level. They are determined uniquely and, in particular, are independent of the demand for commodities. The factors that will change these equilibrium values w^{LR}, n^{LR} and y^{f} are shifts in the production function – with implied shifts in the demand for labor – and in the supply of labor. Other shocks to the economy which do not bring about these shifts will not alter w^{LR}, n^{LR} and y^{f}. Among these shocks are changes in the monetary and fiscal policy variables, which change aggregate demand but do not appear as arguments of the production function and the supply function of labor. This is a very strong implication for the equilibrium states of the neoclassical model. It implies that aggregate demand policies, such as monetary and fiscal policies, cannot affect the equilibrium levels of wages, employment and output in the economy. This implication will be discussed below at greater length.

The irrelevance of aggregate demand for equilibrium output and employment and its empirical validity

Since y^f is independent of the demand side of the economy, the preceding analysis implies that:

1 Aggregate demand and shifts in it cannot change the equilibrium levels of output, employment, real wage and other real variables. Aggregate demand is irrelevant to their determination.

2 Since the model relies on equilibrium for its results, it has the implicit assumption that the economy has adequate and sufficiently fast-reacting equilibrating mechanisms to force aggregate demand y^d into equality with y^f continuously or in a short enough period for the deficit or excess of demand not to affect the production and employment decisions of firms and/or the consumption demand and labor supply decisions of households. Therefore, we might caricature the adjustment process as: "the equilibrium level of supply creates its own demand"[4] through the equilibrating variations in prices, wages and interest rates. This is quite a strong assumption and not all economies in all possible stages of development or of the business cycle meet it.[5]

The irrelevance of aggregate demand and, by implication, of monetary and fiscal policies, which can change that demand, for the determination of output and unemployment is an extremely strong implication of the equilibrium properties of the neoclassical model. Comparison of this implication with the stylized facts given at the beginning of this chapter shows that the implication is clearly not valid. Therefore, in the preceding part of the neoclassical model, either the equilibrium assumption must be abandoned, or its specifications of the production process or of labor demand and labor supply have to be modified. As discussed later, the Lucas supply analysis makes this modification for the production process, while Friedman's expectations-augmented analysis does so for labor demand and supply.

General equilibrium: aggregate demand and supply analysis

Chapter 13 and this chapter have so far specified the markets for commodities, money and labor. The foreign exchange market has been taken to be in equilibrium through appropriate changes in the exchange rate. We have not specified the market for bonds, which are defined as non-monetary financial assets, even though this is one of the four goods in the macroeconomic model. This omission is justified by Walras's law, which specifies that in a four-good economy, if three of the goods markets are in equilibrium the market for the fourth good must also be in equilibrium. Therefore, at the general equilibrium (y^{LR}, r^{LR}, P^{LR}) in the preceding analysis, the bond market will also be in equilibrium and can be omitted from explicit consideration.

A full view of the economy requires simultaneous consideration of all markets, and general equilibrium in the economy implies a simultaneous solution to the equilibrium equations for all the three sectors. We consider this in two alternate ways, demand–supply analysis and the IS–LM analysis.

Demand–supply analysis

The aggregate supply equation derived so far is:

Aggregate supply:

$$y^{s} = y^{LR} = y^{f} \tag{9}$$

The aggregate demand equation is the one derived from either the IS–LM analysis or the IS–IRT analysis of the preceding chapter. Since there are two alternate aggregate demand equations, we use only their general form, stated as:

$$y^{d} = y^{d}(P; g, \theta) \tag{10}$$

where g is the vector of fiscal policy variables and θ is the relevant monetary policy variable. Note that (10) assumes that Ricardian equivalence does not hold. If it does so, then, as shown in the preceding chapter, fiscal variables will not be in the aggregate demand function, so that the AD function will become $y^{d}(P\ \ \theta)$. The validity of Ricardian equivalence is doubtful, so we proceed with the aggregate dem;and function $y^{d}(P\ g, \theta\ )$.

Equilibrium in the commodity market requires that: ;

$$y^{s} = y^{d}(P; g, \theta) \tag{11}$$

(9) and (11) have two endogenous variables: y and P. Of these two equations, (9) clearly determines y, even without reference to (10), as being equal to y^{LR}. Therefore, the aggregate demand equation (10) can only determine P, with y on its left side being set equal to y^{LR}.

Aggregate demand and supply curves

The above conclusion is illustrated in Figure 14.3. Equation (9) implies that the *long-run* aggregate supply curve LAS is vertical while, as shown in the preceding chapter, (10) implies that the open economy aggregate demand curve AD has a negative slope. An examination of this figure clearly shows that shifts in the aggregate demand curve will not change the equilibrium output but only the price level, while changes in the

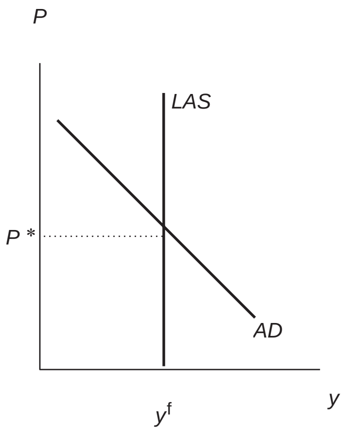


Figure 14.3

aggregate supply will change both output and price level. This is a very strong conclusion and is, as we shall see later in this and the next three chapters, at the heart of the debate on the ineffectiveness of monetary and fiscal policies for the equilibrium states of the neoclassical model and its related modern classical and new classical models.

Equilibrium output and prices

Equations (9) to (11) determine the long-run equilibrium values of output *y* and price level *P* from:

$$y^{\mathrm{LR}} = y^{\mathrm{f}} \qquad (12)$$

$$P = f(g, \theta; y^{\mathrm{LR}}) \qquad (13)$$

so that monetary and fiscal policies can change aggregate demand and the price level, but not long-run output. In the presence of Ricardian equivalence, $P\,f\,(\theta,\ y^{\mathrm{LR}})$, so that changes in monetary policy and in y^{LR} can change the price level, but fiscal policy cannot do so.

In terms of the general equilibrium neoclassical model, (12) determines output and (13) determines the price level. These equations imply the aggregate demand, monetary and fiscal multipliers on output as:

$$\partial y^{\mathrm{LR}}/\partial y^{\mathrm{d}} = 0, \qquad \partial P^{\mathrm{LR}}/\partial y^{\mathrm{d}} > 0 \qquad (14)$$

$$\partial y^{\mathrm{LR}}/\partial g = 0, \qquad \partial P^{\mathrm{LR}}/\partial g \geq 0 \qquad (15)^{6}$$

$$\partial y^{\mathrm{LR}}/\partial \theta = 0, \qquad \partial P^{\mathrm{LR}}/\partial \theta > 0 \qquad (16)$$

which clearly indicate that the long-run equilibrium level of output is not responsive to monetary policies, whereas that of the price level is responsive to these policies.[7] This also applies to fiscal policies, except under Ricardian equivalence when fiscal policy affects neither aggregate demand nor output nor the price level.

Supply shifts

The impact of a change in output on the price level, represented by $\partial P/\partial y^{\mathrm{LR}}$, is negative in both the IS–LM and IS–IRT models, though its magnitude depends on whether the money supply or the interest rate is the exogenous monetary policy variable. The basic reason for the negative impact of output increases on the price level is that an increase in output raises the transactions demand for money, which has to be offset by lower prices. We leave it to the interested reader to derive $\partial P/\partial y$ for the IS–LM and IS–IRT models, using the information given in this and the last chapter.

Iterative structure of the neoclassical model

Another procedure for studying simultaneous equilibrium in all sectors of the economy emerges if the final *equilibrium* equation for each of the sectors is examined separately in the overall problem. The following equations (17) to (22) incorporate information on all the sectors.

Production–employment sector:

$$y = y^{\mathrm{f}} \tag{17}$$

The commodity market IS equation:

$$y = y(r, P; g, \theta) = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y)}\right) \left\{ c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}} + \frac{1}{\rho^{\mathrm{r}}}\left(-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\right)\right\} \tag{18}$$

Fisher equation:

$$R = (1 + r)(1 + \pi^{\mathrm{e}}) \tag{19}$$

From the employment–output sector equilibrium, we know that the long-run equilibrium output is y^{f}. Substitute this equilibrium level of output into the IS relationship (18), which yields the long-run equilibrium real rate of interest as r^{LR}_0:

$$r^{\mathrm{LR}}_0 = \frac{1}{i_r}\left\{ c_0 - c_y t_0 + i_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}} + \frac{1}{\rho^{\mathrm{r}}}\left(-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\right)\right\} - \frac{\left(1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{\mathrm{cy}}(1 - t_y)\right) y^{\mathrm{f}}}{i_{\mathrm{r}}} \tag{20}$$

We now know the long-run equilibrium level of the real interest rate, even without introducing the monetary sector into the analysis so far.

Money market equilibrium depends on whether the central bank uses the interest rate or the money supply as the exogenous monetary policy variable. The relevant equations are:

LM equation in the IS–LM model:

$$\overline{M}/P = m_y y + (FW_0 - m_R R) \tag{21}$$

Money market equilibrium condition in the IS–IRT model:

$$M/P = m_y y + (FW_0 - m_R \overline{R}) \tag{22}$$

The following analysis deals separately with each of these cases.

Determination of the price level for an exogenously given money supply (the IS–LM model)

The preceding results imply that the quantity of the money supply is irrelevant to the determination of the long-run equilibrium values of both output and the real rate of interest. To determine the price level in the neoclassical model based on the IS–LM analysis, we start with the Fisher equation, which determines the nominal interest rate from $(1 + r^{\mathrm{LR}}_0)(1 + \pi^{\mathrm{e}})$, so that, depending on the expected inflation rate, different nominal interest rates are consistent with r^{LR}_0.

In the LM equation with an exogenously given money supply, substituting the long-run equilibrium level of income y^{f} and the nominal interest rate from the Fisher equation as approximately equal to $(r^{\mathrm{LR}}_0 + \pi^{\mathrm{e}})$ yields the equilibrium price level as:

$$P = \frac{\alpha \cdot i_r \cdot M_0}{m_R\left\{y^{\mathrm{f}} - \alpha \cdot \left[c_0 - c_y t_0 + i_0 - \frac{i_r}{m_R} FW_0 + i_r \pi^{\mathrm{e}} + g_0 + x_{c0} - x_c\,\rho^{\mathrm{r}} + \frac{1}{\rho^{\mathrm{r}}}\left(-z_{c0} + z_{cy} t_0 - z_{c\rho}\,\rho^{\mathrm{r}}\right)\right]\right\}} \tag{23}$$

where:

$$\alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y) + i_r \frac{m_y}{m_R}}\right)$$

Note the sequence in the above procedure. The production–employment sector alone determines the full-employment output in (17), without any reference to the interest rate and the price level; the expenditure sector then determines the long-run equilibrium real interest rate, without any reference to the price level or the monetary sector; but the price level is determined by reference to all the sectors of the economy.

In the special quantity theory case when $m_R = 0$, the monetary sector condition with $y = y^{\mathrm{f}}$ simplifies to:

$$M/P = m_y y^{\mathrm{f}}$$

which can be rearranged as:

$$P = \left(\frac{1}{m_y y^{\mathrm{f}}}\right) M \tag{24}$$

which does not include either *r* or *R* as a variable, so that we need to know only the money supply and output to determine the price level. This can be taken to be a modern derivation of the quantity theory, as in Pigou's version in Chapter 2. In it, the dependence of the price level upon the interest rate is eliminated by the assumption that the interest sensitivity of money demand is zero and the economy has full-employment output.

Determination of the price level for an exogenously given interest rate (the IS–IRT model)

In the IS–IRT model of the preceding chapter, it was assumed that the central bank sets the real interest rate, with its level specified as $r^T{}_0$. Aggregate demand y^d is then given by the commodity market at the given real interest rate. This AD equation is:

$$y^d = y(r^T_0, P; g, \theta) = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)}\right) \cdot \left\{ \left[c_0 - c_y t_0 + i_0 - i_r r^T_0 + g + x_{c0} - x_{c\rho}\rho^r\right] + \frac{1}{\rho^r}\left[-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^r\right]\right\} \tag{25}$$

Given output supply at y^f, and the real interest rate at r^T, this equation determines the price level P.

Policy implications: the ineffectiveness of monetary and fiscal policies in changing output and employment

Important policy implications follow from the iterative nature in the long-run equilibrium of the neoclassical model. No matter what is the monetary policy variable, neither monetary nor fiscal policies can affect equilibrium output, since neither appears in (17). Therefore, these policies are useless for increasing the long-run equilibrium levels of employment and reducing the equilibrium level of unemployment. However, with the economy in long-run equilibrium, these policies are not only useless, they are also not needed since the long-run equilibrium employment is at full employment. Hence, in the long run, there is *no need or scope* for the authorities to pursue policies to increase employment or output, which are both at their full-employment levels. Any such attempt will be ineffective.

If we compare these implications on the irrelevance of monetary and fiscal policies for output and employment with the stylized facts, these implications are clearly invalid. Expansionary monetary and fiscal policies do increase aggregate output and lower unemployment, and the reverse is true for contractionary policies. Therefore, the preceding model needs to be modified in order to be valid and useful. The classical economists do this, as explained later, by introducing uncertainty, with errors or misperceptions in expectations, into the model.

The rate of unemployment and the natural rate of unemployment

The level of unemployment is defined as:

$$U = L - n \tag{26}$$

where:

$U =$ level of unemployment
$L =$ labor force.

Since $n \leq L$, unemployment is always non-negative. Assuming L to be exogenously given as $\underline{L}$, so that it does not vary with the real wage, the labor force will be the sum total of workers who are able and willing to work at *any* wage. In this case, L represents the maximum amount of potential employment in the economy. But if $L = L(w)$, it is likely that the number of workers willing to work increases as real wages rise, so that $L^J \geq 0$. For our fairly basic analysis at this point, we will make the former assumption.

The natural rate of unemployment

The *long-run equilibrium level* of unemployment U^{LR} is:

$$U^{\mathrm{LR}} = L - n^{\mathrm{LR}} \tag{27}$$

The *long-run equilibrium rate* of unemployment u^{n}, with the superscript n standing for "natural," which itself stands for long-run equilibrium, is:

$$u^{\mathrm{n}} = U^{\mathrm{LR}}/L = 1 - n^{\mathrm{LR}}/L \tag{28}$$

From (17), since aggregate demand and its determinants cannot change output, they also cannot change unemployment. Hence, from (28),

$$\partial u^{\mathrm{n}}/\partial y^{\mathrm{d}} = \partial u^{\mathrm{n}}/\partial \theta = \partial u^{\mathrm{n}}/\partial g = 0$$

where θ is the monetary policy variable and g is the fiscal variable. Hence, in the neoclassical model, since n^{LR} is independent of the demand side of the economy and L is exogenous, u^{n} is also independent of changes in aggregate demand and, therefore, of monetary and fiscal policies. Note that $u^{\mathrm{n}} > 0$ by virtue of structural, frictional, search and seasonal unemployment in the economy, which prevent the employment of all members of the labor force since some would have inappropriate skills and education, be in inappropriate locations or require wages in excess of their marginal productivity in the current state of the economy.

On the characteristics of the natural rate of unemployment in the neoclassical model, this rate cannot be changed by monetary or fiscal policies (Friedman, 1977). It does, however, depend upon the supply structure – the labor market relationships and the production function – of the economy and will change as the supply structure changes. Technical change and changes in educational and skill requirements, the level of education of the labor force, the availability of information on jobs and workers, the location of industry, etc., are thus likely to change the natural rate of unemployment. This rate is therefore itself a variable, though not one that can be changed by demand shifts in the economy, including the pursuit of monetary and fiscal policies. Shifts in the supply side of the economy can change the natural rate. Among these shifts is technical change and shifts in the industrial structure due, among other things, to shifts in the structure of demand among the sectors of the economy.

The natural rate of unemployment rises during a transition from one industrial–agricultural structure of the economy to a different one. Suppose that industry A is declining and laying off workers while industry B is expanding its labor force. The process of transfer of workers involves searching for new jobs by the laid-off workers, so that search unemployment increases during the transition. Further, some of the laid-off workers may possess skills not needed in industry B and may become permanently unemployed. This increases structural unemployment in the economy. That is, the shift in the economy's industrial structure induces a *transitional* increase in the natural rate of unemployment, but it can also imply a long-run shift in that rate.

IS–LM version of the neoclassical model in a diagrammatic form

Figure 14.3 brings together the information in equations (17) to (22). (18) to (22) specify the downward-sloping AD curve. From (17), since output does not depend on *P*, the aggregate supply (AS) curve is vertical at y^f. It is often called the "full-employment (y^f) curve." Equilibrium in all the sectors of the economy exists at the point (P^*, y^f).

An expansionary monetary policy with an exogenous money supply (the IS–LM model)

Figure 14.4 illustrates the IS–LM model in the (r, y) space. Equation (17) specifies the LAS curve at the full-employment output y^f, (18) specifies the IS curve, and (19) and (21) are used to specify the LM curve. The intersection of the IS and LM curves determines the level of aggregate demand. General equilibrium in the commodity, money and output sectors requires that all three curves intersect at the same point. This is the case shown in Figure 14.4.

We next provide some examples of how the IS–LM analysis and diagram can be manipulated for comparative static studies. This is done in Figure 14.5 for monetary policy. Suppose that the economy is initially in overall equilibrium at the point a and the money supply increases, shifting the LM curve from LM_0 to LM_1. The new equilibrium between the monetary and expenditure sectors is shown by the point *d* and represents nominal aggregate demand. But output specified by the output–employment sector is at y^f. Since aggregate demand at *d* exceeds the supply of output y^f, prices rise. As *P* rises, the LM curve shifts towards the left. However, the rise in *P* leaves the IS and AS curves unchanged. Prices will then continue to rise as long as aggregate demand for output exceeds its supply. This will occur until the leftward shifts in the LM curve due to the price increases take it sufficiently back (i.e. from LM_1 to LM_1') to pass through the initial equilibrium point a. In short, in terms of equilibrium states, an increase in the money supply increases

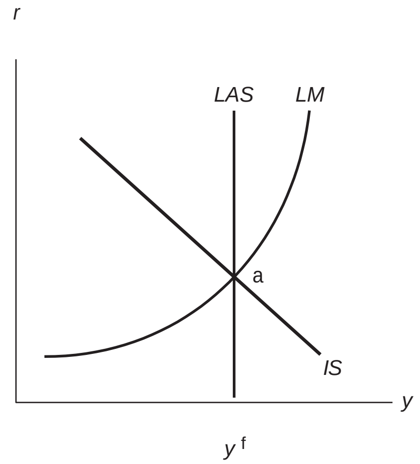


Figure 14.4

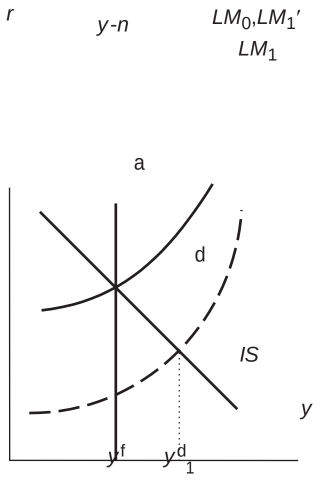


Figure 14.5

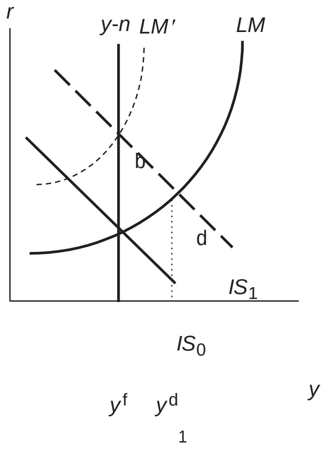


Figure 14.6

aggregate demand, without affecting output, and raises prices to a new level, with the increase in the price level being proportionate to the increase in the money supply. The interest rate is unchanged between the old and the new equilibrium positions. Comparing these implications with the stylized facts given at the beginning of this chapter shows that they are not valid for the short run. Therefore, the model presented so far needs to be modified.

Fundamental assumptions of the Walrasian equilibrium analysis

The preceding analysis has focused on the long-run equilibrium states of the model and is a succinct macroeconomic version of the *Walrasian model in the absence of uncertainty*. This equilibrium analysis makes four fundamental assumptions, plus a fifth one that specifies that the findings are for the certainty case. These are:

(I) *Flexible prices and wages, and the stability of markets*
The prices of all the goods in the economy are assumed to be flexible and adjust to equate demand and supply in the relevant market. They increase if there is excess demand and decrease if there is excess supply. These prices include wages, which is the price of labor. Wages are flexible in both nominal and real terms.

(II) *Perfect market hypothesis*[8]
Each market has perfect competition and clears *continuously*, so that we can focus on the study of competitive general equilibrium in the economy and its properties, while largely ignoring the disequilibrium values of the variables.

(III) *Transparency of equilibrium prices*
All agents, in making their demand and supply decisions, assume that such market clearance will occur instantly after any disturbance and know or anticipate (or are informed by an agency such as a "Walrasian auctioneer" or "market coordinator") the prices at which it will occur. Further, all agents plan to produce, consume, demand money and supply labor at *only* these equilibrium prices.[9]

(IV) *Notional demand and supply functions* All economic agents assume that they will be able to buy or sell as much as they want to at the long-run equilibrium prices. The demand and supply functions derived under this assumption are known as *notional* (as against effective) *demand and supply functions*.

The Walrasian-based models now also include:

(V) *Assumptions on uncertainty*
If the model assumes certainty, its results provide the long-run equilibrium of the economy. Its extension to include uncertainty will produce deviations from this long-run equilibrium, with the nature of the deviations depending on the way uncertainty is handled in the extended model. When the equilibrium of the economy under uncertainty differs from the long-run equilibrium of the model, such an equilibrium is designated as a short-run equilibrium. One of the causes of such a deviation from the long-run equilibrium is errors in price expectations. However, there can also be many other causes of such deviations.

The current versions of the Walrasian-based macroeconomic models assume the rational expectations hypothesis, with the long-run general equilibrium outcomes used to specify the rationally expected values of the relevant variables.

The classical equilibrium analysis of the economy and its policy recommendations require the applicability of the above fundamental assumptions: flexible prices and wages, continuous market clearance, transparency of equilibrium prices and notional demand and supply functions. Each assumption is related to the others but is still distinct. Any one or more of these assumptions may not be relevant to or valid for a particular stage or at a particular time in an economy.

As shown above, monetary policy is neutral in the long-run equilibrium states of the neoclassical model since changes in monetary policy affect only the nominal but not the real variables. Hence, the neoclassical model implies that, in its long-run equilibrium, monetary policy cannot change output and employment in the economy. In fact, with the economy at full employment, there is also no need for the pursuit of monetary policy. Conversely, if the above assumptions hold, changes in the money supply also cannot be detrimental for long-run output and employment in the economy. In particular, decreases in the money supply will not decrease output and employment and force the economy into a recession. Monetary policy is benign (harmless) in such a context. Note that these are assertions about the long-run analytical state, not about the short-run equilibrium or disequilibrium.

Disequilibrium in the neoclassical model and thenon-neutrality of money

Note that the neoclassical model does not assert that its long-run equilibrium must always exist – as if it were an identity – and therefore it allows the possibility that the economy can be sometimes in short-run equilibrium or in disequilibrium. Further, for the study of disequilibrium to be a potentially useful exercise requires the belief that the economy will be away from its full-employment equilibrium state for significant periods of time. Intuitively, continuous general equilibrium requires, for example, the belief that an increase in the money supply *immediately* causes a proportionate increase in the price level and that a decrease in the money supply does not cause a recession in output and employment. There is considerable evidence to show that these requirements are not always, or most of the time, met in most real-world economies. Nor did the major proponents in the traditional classical and neoclassical tradition claim that they did. Among those who allowed for the existence of disequilibrium for significant periods of time, and the impact of money supply changes on output (i.e. the non-neutrality of money) during disequilibrium, were Hume, Marshall, Fisher and Pigou in the traditional classical school, and Friedman and the St Louis monetarists in the neoclassical one.

However, when the classical economists allowed for disequilibrium, they maintained that any such state of disequilibrium incorporates certain forces brought about by markets (not firms) that will force the economy into equilibrium. Among these forces are price changes and the Pigou effect, including the real balance effect. These were touched upon in Chapter 3 and will be explored more fully in Chapter 18. We explain them again, though briefly, in the following subsection as a reminder and for completeness of the neoclassical model.

Pigou and real balance effects

The *Pigou effect* is associated with the contributions of A.C. Pigou (at Cambridge University in England in the first half of the twentieth century) in the debate between the traditional classical school and Keynes in the 1930s. The Pigou effect is another name for the effect of real wealth (defined as including all financial assets) on (real) consumption resulting from a change in the price level. That is, the Pigou effect is represented by:

$$\text{Pigou effect} = \underset{+}{[\partial c/\partial \text{ real wealth}]} \cdot \underset{-}{[\partial \text{ real wealth}/\partial P]} < 0$$

The Pigou effect works in the following manner. A disequilibrium with deficient demand for commodities will cause a fall in their prices. Since the household's wealth includes financial assets, this fall in the price level will increase the household's wealth, which, in turn, will cause consumption to rise. The latter will bring about an increase in aggregate demand in the economy. This process will continue until the demand deficiency is eliminated – that is, until the economy returns to equilibrium.[10]

The *real balance effect* is associated with the contributions of Don Patinkin during the 1940s to the 1960s and represented a refinement of the neoclassical model for the analysis of disequilibrium (Patinkin, 1965). This effect represents the impact of changes in the price level on consumption through changes in the real value of money holdings. It works as follows. A price fall due to a demand deficiency will increase the real value of money holdings and thereby increase the household's wealth. This will lead to an increase in consumption and therefore in aggregate demand. The real balance effect will continue until the demand deficiency and its associated price level decreases are eliminated.

That is, the real balance effect is represented by:

$$\text{Real balance effect} = \underset{+}{[\partial c/\partial\,(M/P]} \cdot \underset{-}{[\partial(M/P))/\partial P]} < 0$$

Hence, the real balance effect and the Pigou effect are equilibrating mechanisms of the neoclassical model and require flexible prices. They were shown to imply that a fall in aggregate demand would produce a fall in prices, which would increase aggregate demand because of an increase in the real value of financial assets in the case of the Pigou effect and of money balances in the case of the Patinkin effect. Their *analytical* relevance, though under the *ceteris paribus* clause, is significant and beyond dispute. However,

> [Pigou himself] described what later came to be called the "The Pigou Effect" as a mere toy, based on "so extremely improbable assumptions" as to never be played "on the checkerboard of real life". … It would not work in any except the most formal version of a most naïve model [because, outside its *ceteris paribus* assumptions, a fall in aggregate demand would also bring about]… simultaneous bankruptcies and deflation [which] keep shifting both the LM and IS functions, and therefore the aggregate demand function towards the origin. The result is much more likely to be a depression rather than full employment.
>
> (Pesek, 1988, pp. 6–7).

Further, the analyses of the real balance and the Pigou effects do not, a priori, provide any guidance on the time the neoclassical economy will need to return to long-run equilibrium under their impetus. In particular, the real balance effect can be quite weak, so that the neoclassical economy could react to an exogenous fall in demand by a very slow movement towards long-run equilibrium and, therefore, remain away from it for a significant period. Hence, it is important to analyze the short run and disequilibrium properties of the neoclassical model and derive its policy implications.

Macroeconomic theory uses two main channels for the effects of money supply changes on aggregate demand. One of these is because changes in the money supply change wealth and real balances, which changes consumption expenditures (the direct transmission channel). The other is through the effect of money supply changes on interest rates, which changes investment expenditures (the indirect transmission channel). The direct effects are considered to be relatively small over the business cycle, so that most macroeconomic models, including the popular IS–LM one, ignore them altogether.[11] Consequently, these models embody only the indirect transmission channel of monetary policy effects through interest rates.

Causes of deviations from long-run equilibrium

The actual economy may not be in or close to its long-run equilibrium because of:

1 Errors in expectations, either in the commodity or/and in labor markets. The analysis of this scenario is presented in the next section on the short-run equilibrium of the neoclassical model.

2 Costs of adjusting prices, wages, employment and output. The analyses of these various scenarios are presented in the next chapter on the Keynesian paradigm.

3 The absence of a mechanism for instantly restoring equilibrium. Note that the assumption of perfect competition does not a priori specify the chronological time needed by the “invisible hand of competition” to return an economy, following a shock, to its full-employment equilibrium.[12] Along with the absence of a mechanism for instantly restoring the full-employment equilibrium, the competitive economy also does not have an instantaneous mechanism operating in disequilibrium for computing the new price level and informing all firms and households of the new prices of the products. Further, there is no guarantee given to the firms that they could sell all they wanted at these prices, nor is there any guarantee that the workers will receive or recoup the incomes lost while they were unemployed. Combining this lack of a guarantee with the plausible possibility that firms and households may respond faster than markets to disequilibrium and will do so on the basis of expectations of the quantity demanded and jobs available, could produce a disequilibrium path that keeps the economy out of equilibrium for quite some (chronological) time. This scenario is sketched in the next chapter.

4 Firms are in monopolistic competition with sticky prices (see Chapter 15).

The relationship between the money supply and the price level: the heritage of ideas

The basic comparative static conclusions of the neoclassical macroeconomic model of this chapter were presented several centuries ago. The following quote illustrates them from the writings of David Hume, one of the founders of classical economics.

> Money is nothing but the representation of labor and commodities, and serves only as a method of rating or estimating them. Where coin is in greater plenty – as a greater quantity of it is required to represent the same quantity of goods – it can have no effect, either good or bad, taking a nation within itself; any more than it would make an alteration in a merchant's books, if instead of the Arabian method of notation, which requires few characters, he should make use of the Roman, which requires a great many.
>
> (Hume, *Of Money*, 1752).

This quote is an assertion of the basic proposition of the quantity theory of money: an increase in the money supply causes a proportional increase in prices. Further, Hume asserts that changes in the supply of money do not change real output in the economy but correspond to a change in the unit of account. The quantity theory was presented in Chapter 2.

The conclusions of this theory on the proportionate relationship between the money stock and prices, and the inability of the monetary and fiscal authorities to control output and employment, apply in equilibrium. They do not necessarily apply in disequilibrium – that is, during the adjustment from one equilibrium to another. In fact, many supporters of this theory, in Hume's time and down to the present, have viewed changes in the money stock as exerting a powerful influence on output, employment and other variables in the adjustment process. Hume himself described this process as follows.

> Notwithstanding this conclusion, which must be allowed just, it is certain that since the discovery of the mines in America, industry has increased in all the nations of Europe, except in the possessors of those mines; and this may justly be ascribed, among other reasons, to the increase of gold and silver. Accordingly, we find that in every kingdom, into which money begins to flow in greater abundance than formerly, everything takes on a new face; labor and industry gain life; the merchant becomes more enterprising, and even the farmer follows his plough with greater alacrity and attention…
>
> To account then for this phenomenon, we must consider, that though the high price of commodities be a necessary consequence of the increase of gold and silver, yet it follows not immediately upon that increase; but some time is required before the money circulates through the whole state, and makes its effect be felt on all ranks of people. At first, no alteration is perceived; by degrees the price rises, first of one commodity, then of another; till the whole at last reaches a just proportion with the new quantity of specie which is in the kingdom. In my opinion, it is only in this interval or intermediate situation, between the acquisition of money and the rise of prices, that the increasing quantity of gold and silver is favorable to industry.
>
> (Hume, *Of Money*, 1752).

Hume's opinions on the disequilibrium path of adjustment serve as a note of caution against total reliance on the comparative static results of the neoclassical model and against the belief that the pre-1936 classical economists *assumed* that the economy always functions in full employment. Hume's analysis of disequilibrium shows that the economy can be in disequilibrium for some time and that money will not be neutral during this period.

Almost two centuries later, Pigou, a twentieth-century economist in the classical tradition, expressed similar ideas in the following excerpts from his book, *Money, A Veil* (1941).

> Money – the institution of money – is an extremely valuable social instrument, making a large contribution to economic welfare. … if there were no generally accepted money, many of these transactions would not be worth undertaking, and as a direct consequence the division of labor would be hampered and less services and goods would be produced. Thus, not only would real income be allocated less satisfactorily, from the standpoint of economic welfare, among different sorts of goods, but it would also contain smaller amounts of many, if not of all sorts. … Obviously then money is not merely a veil or a garment or a wrapper. Like the laws of property and contract, it constitutes at the least a very useful lubricant, enabling the economic machine to function continuously and smoothly. …
>
> So far everyone would be agreed. But now an important distinction must be drawn. The institution of money is, as we have seen, a powerful instrument promoting wealth and welfare. But the number of units of money embodied in that instrument is, in general, of no significance. It is all one whether the garment, or the veil, is thick or thin. I do not mean, of course, that it is immaterial whether the number of units of money is held constant, or is variable in one manner, or is variable in another manner in relation to other economic happenings. I mean that if, other things being equal, over a series of months or years the stock of money contains successively $m \times 1$, $m \times 2$, $m \times 3$. … units, it makes no difference what the value of *m* is. A doubled value of *m* throughout means simply doubled prices throughout of every type of goods – subject, of course, to the rate of interest not being reckoned for this purpose as a price – and all real happenings are exactly what they would have been with a value of *m* half as large. The reason for this is that, money being only useful because it exchanges for other things, a larger quantity does not, as with other things, carry more satisfaction on its back than a smaller quantity, but the same satisfaction.
>
> (Pigou, 1941, Ch. 4).

In this quote, Pigou is clearly expressing the long-run equilibrium results following an increase in the money supply. Note that what is missing in this story are Hume's conclusions on the impact of changes in the money supply during the ensuing adjustment period. However, as his other writings show, Pigou was quite aware of such an adjustment period and of the impact of money supply variations in causing fluctuations in employment and output.

The classical and neoclassical tradition, economic liberalism and laissez faire

The long-run equilibrium analysis of the neoclassical model in this chapter implies that the economy functions at full employment, with full-employment output, so that there is no scope for monetary and fiscal demand management policies for such an economy. Such a viewpoint is part of the classical philosophy of economic liberalism, which can be broadly formulated as stating that the economy performs at its best by itself and that the state cannot improve on its performance. This is usually also supplemented by the proposition that any intervention by the state, even with the intention of improving upon the performance of the economy, worsens its performance. These propositions imply that the goods and input markets should be free and that free enterprise should be the desired standard. However, market imperfections such as imperfect competition, oligopoly, monopoly or monopsony, externalities, etc., could and often do exist in the actual economy. Advocates of the *strong form of economic liberalism* argue that, even in such cases, the economy should be left as it is and the state should not attempt to eliminate such imperfections; the imperfections are minor and, even when they are not of minor significance, there is no guarantee that state intervention will achieve a net improvement since its intervention might eliminate some imperfections while introducing others. A *weaker version of economic liberalism* allows the state to intervene to eliminate market imperfections through selective policies, though without assigning a role to general monetary and fiscal policies.

To be credible, the general liberalism philosophy has its underpinnings in the nation's political, economic and social ideology, and in the public's perception and goals for the actual economic and social performance of the nation. In its general approach, the underlying philosophical basis of liberalism was provided by the utilitarianism approach of Jeremy Bentham and his followers in the first half of the nineteenth century. The main tenet of this approach was that economic agents (households and firms), working in their own best interests (utility and profit maximization), would ensure that social welfare was maximized.[13] Consequently, the economy should be left alone by the government and regulatory agencies. This policy approach was summarized in the term *laissez faire*. In its economic aspects, the liberalism philosophy needed a theoretical economic model that could justify its economic policy recommendations. This model was provided at the macroeconomic level by the traditional classical approach in the pre-Keynesian period and is currently the neoclassical one – with the modern and new classical models among its versions.

The economic and social problems of nineteenth-century Britain, with rapid industrialization and urbanization, were sufficiently acute and transparent to lead to a gradual evolution of political and economic thought away from liberalism and *laissez faire* and towards some form of socialism, with support for some degree of state intervention in the economy. This evolution of ideas was widespread during the latter half of the nineteenth century and early twentieth century. The Great Depression of the 1930s destroyed the public's and economists' faith in *laissez faire*, so that Keynes's publication of *The General Theory* in 1936, with its encouragement to the state to use monetary and fiscal policies to improve on a poorly performing economy, proved to be timely and readily won acceptance from most economists and the public. Economic liberalism was eclipsed by Keynesianism for the decades from the 1930s to the 1970s. The Keynesian approach is the subject of the next chapter.

In economics, the traditional classical ideas were reformulated and rebottled in the form of the neoclassical theory during the decades of the 1940s to the 1970s. Since the 1970s, these ideas, in the form of the modern classical model and with the agenda of making microeconomics the foundation of macroeconomics, have again become the dominant approach in macroeconomics. Their return to this dominance detoured briefly through Monetarism during the 1970s. They are currently supported by the new classical approach developed during the 1970s and the 1980s.

11 .

Some major misconceptions about traditional classical andneoclassical approaches

A common misconception nowadays is that the traditional classical and neoclassical economists believed that the economy functioned well enough to maintain full employment most of the time or that it had a fast tendency to return to full employment following a disturbance and a decline in employment. In fact, many believed that "the economic system is essentially unstable" (Patinkin, 1969, p. 50).[14] Another misconception nowadays is that the classical and neoclassical economists believed that money was neutral in practice and in theory. In fact, as the business cycle literature of the nineteenth and early twentieth centuries amply shows, it was a common and strongly held belief that fluctuations in the money supply were a major cause of recessions, with declines in employment and output, and of booms in real economic activity.

During the period of traditional classical dominance (mid-eighteenth century to 1936), booms and recessions were common and sometimes quite severe. It was a common observation among economists that the velocity of circulation of money did change, and did so over booms and recessions. In fact, many economists believed that there could occur – and did occur – "extreme alternations of hoarding and dishoarding" (p. 50) because of changes in expectations, and that the changes in the money supply and in its velocity were major sources of business fluctuations, as attested by Don Patinkin (1972). Further, many economists believed that these fluctuations were:

> exacerbated by the "perverse" behavior of the banking system, which expands credit in booms and contracts it in depressions.
>
> (Patinkin, 1969, p. 51).

Among the reasons for the real effects of money supply and velocity changes was that:

> Costs have a tendency to move more slowly than do the more flexible selling prices [i.e. firms' costs were sticky relative to their final prices]
>
> (Patinkin, 1969, p. 57).

> Sticky prices are peculiarly resistant to downward pressure. … [To sum up, it was generally recognized that] cycles and depressions [are] an inherent feature of "capitalism." Such a system must use money, and the circulation of money is not a phenomenon which naturally tends to establish and maintain an equilibrium level. Its equilibrium is vague and highly unstable.
>
> (Patinkin, 1969, pp. 63–64).

In view of such strong effects of money supply variations on employment and output, what the traditional classical economists believed was that money was neutral in the long run, but not in the short run or over the business cycle. Its neutrality in the long run was often less a matter of analysis than of belief, which was sometimes reflected in the proposition known as Say’s law (see Chapter 18).

On policy issues, in view of the reasons given above, the traditional classical school believed that:

> The government has an obligation to undertake a contracyclical policy. The guiding principle of this policy is to change M so as to offset changes in V, and thus generate the full employment level of aggregate demand MV.
>
> Once a deflation has gotten under way, in large modern economy, there is no significant limit which the decline of prices cannot exceed, if the central government fails to use its fiscal powers generously and deliberately.
>
> (Patinkin, 1969, pp. 51, 63).

As the preceding quotes from Patinkin, a foremost neoclassical economist, convincingly show, monetary policy was often envisaged and recommended as a stabilization tool. Fiscal policy was sometimes, but not commonly, considered a possibility since pre-1936 economic analysis did not have its analytical basis nor did it have a theory of the aggregate demand for commodities. The analytical basis for fiscal policy and its recommendation as a major tool for stabilization of aggregate demand were due to Keynes and the Keynesians. As a counter-reformation, Barro’s Ricardian equivalence theorem (Barro, 1974; see also Chapter 13, Section 13.7) sought to again remove fiscal policy from the set of potential stabilization tools.

Uncertainty and expectations in the classical paradigm

The analysis of the neoclassical model so far shows that many of its implications are contradicted by the stylized facts. In particular, monetary and fiscal policies do change output and unemployment, as against the model’s implication that they do not. Therefore, the model has to be modified. Classical economists do so by introducing uncertainty into the model and relying on errors in prices expectations. The two models in this stream are the Friedman model, which relies upon wage contracts and errors in price expectations in labor markets, and the Lucas model, which relies on expectational errors by firms in commodity markets. The following analyses deal with the nature of uncertainty and such errors.

The nature of risk, uncertainty and expectations in economics

Events whose outcomes are not known at the moment of decision making used to be classified in economics in the first half of the twentieth century into risky ones or uncertain ones. The difference between these terms was that an event involves risk if the objective probabilities of its outcomes exist and are known, whereas an event involves uncertainty if the objective probabilities of its outcomes do not exist or are not known. Because of the nature of economic events and/or because of the pervasive imperfection of knowledge involving future outcomes, few economic decisions involve known objective probabilities, so that the standard case in economics is one of uncertainty. However, probability theory finds it unmanageable to include many of the elements of uncertainty such as the vagueness, inadequacy and imperfection of information that distinguish uncertainty from risk. As a consequence, neoclassical economics often abandons the above distinction between risk and uncertainty, and treats the latter as if it were really a case of risk, while Keynesian, especially post-Keynesian, economics often makes a strong distinction between them. For consistency with the literature relevant to this chapter, we shall use the word *uncertainty* as if it were synonymous with *risk*.

Note that for uncertain events, it can be validly postulated that the individual forms subjective probabilities of the anticipated outcomes, with such probabilities being based on whatever knowledge the individual possesses or considers profitable to acquire. Such knowledge can be highly inadequate and imperfect and even the range of anticipated outcomes can differ from the possible ones, so that the subjective probabilities held can be highly erroneous or volatile[15] and would also differ among individuals. In general, classical macroeconomics ignores these problems with subjective probabilities.

Expectations hypotheses in macroeconomics

The two major hypotheses in economics for constructing the expected value of a variable are the adaptive expectations hypothesis and the rational expectations hypothesis (REH). The former is a statistical procedure, while the latter represents an economic theory of expectations. These approaches were used in Chapter 8 to estimate expected and permanent income for the demand for money function. The hypothesis used in this chapter is that of rational expectations, estimating the expected rate of inflation, the expected money supply or the expected level of aggregate demand. The material in Chapter 8 on the rational expectations hypothesis should be reviewed at this stage.

Since the classical macroeconomic models assume that the economy will either stay in equilibrium or soon revert to it, the rationally expected levels of output, unemployment and prices, as of all other endogenous variables, are their long-run equilibrium (full-employment) levels. Their values can therefore be derived from the long-run solution of the model. While this might be acceptable for analytical models that have their focus on the equilibrium values of the variables, the practice of monetary policy requires forecasts of the actual values of these variables at the time the policy will impact on them. Given the long lags in monetary policy, the relevant expectations even by the policy makers are often erroneous. From the perspective of nominal wage contracts, the economic agents must be able to forecast the price level during the duration of the contract. Such forecasts often have errors, so that the real wage received usually differs from the real wage expected by firms and workers to accrue from the contract. We next investigate the impact of expectations on wages, employment and output.

Expectations and the labor market: the expectations-augmented Phillips curve

Output and employment in the context of nominal wage contracts

In industrialized economies, the nominal wage between a firm and its workers is established – whether explicitly negotiated or arrived at by implicit arrangement – ahead of the production and employment decisions by the firm and before the actual price level is known. In arriving at the contracted nominal wage, firms and workers must base their agreement on the *expected* real wage – that is, the nominal wage divided by the expected price level – rather than the a priori unknown *actual* real wage which will apply at the time of employment and production. However, firms can continue to adjust their employment as the price level changes, so that their decisions on employment, as determined by their demand function for labor, will depend on the actual real wage – equal to the established nominal wage divided by the actual price level. This section modifies the preceding neoclassical labor market analyses to incorporate these ideas.

To start, consider the household utility maximization, as in Chapter 3, but now with the addition of uncertainty about the future price level and with labor forming expectations on it. Let the household's expected price level for the period ahead be P^{eh}. Utility maximization would then imply that the supply function of labor is:

$$n^s = n^s(w^{eh})$$

$$= n^s(W/P^{eh}) \quad n^{s\prime} > 0 \tag{29}$$

where:

n^s = labor supply function
w^{eh} = expected real wage, as expected by labor
W = actual nominal wage
P^{eh} = price level expected by households or workers for the duration of the labor contract

Designate the inverse of $n^s(.)$ as $h(n)$, so that inverting (29) and rearranging yields:

$$W^d = P^{eh} \cdot h(n^s) \quad h' > 0 \tag{30}$$

where W^d is the wage demanded by workers in negotiations. Designate the representative firm producing the *i* th product – and being in the *i* th market – as the *i* th firm. Prior to the production period, the profit-maximizing *i* th firm in perfect competition would equate its marginal product of labor to the expected real wage measured in terms of the firm's expected product price, so that:

$$n^d_i = n^d_i(w^{ef}_i)$$

$$= n^d_i(W/p^{ef}_i) \quad n^{d\prime}_i < 0 \tag{31}$$

where:

w^{ef}_i = expected real wage, based on the *i* th firm's expectations of its product price
p^{ef}_i = expected product price, as expected by the *i* th firm.

Aggregating over all firms, let P^{ef} be the average expected price for all firms and n^d the aggregate demand for labor. The aggregate demand for labor is given by:

$$n^d = n^d(W/P^{ef}) \quad n^{d\prime} < 0 \tag{32}$$

Designating the inverse of $n^d(.)$ as $f(n^d)$, the nominal wage W^o offered in the wage negotiations by the firms is:

$$W^{o} = P^{ef} \cdot f(n^{d}) \quad f^{J} < 0 \tag{33}$$

Assuming that the market-clearing (that is, with n^{d} n^{s} n) nominal wage is negotiated, the wage process based on (30) and (33) would yield the equilibrium nominal wage W^{c} as:

$$W^{\mathrm{c}} = P^{\mathrm{ef}} \cdot h(n) = P^{\mathrm{ef}} \cdot f(n) \qquad (33^{\mathrm{J}})$$

where the superscript c indicates the contractual nature of the wage.[16] W^{c} can be obtained by directly solving (29) and (32) and has the general functional form:

$$W^{\mathrm{c}} = g(P^{\mathrm{ef}}, P^{\mathrm{eh}}) \quad \partial g/\partial P^{\mathrm{ef}}, \partial g/\partial P^{\mathrm{eh}} > 0 \qquad (34)$$

The explanation for the signs of the derivatives is as follows. An increase in the firm's expected price level increases its willingness to agree to higher nominal wages, and an increase in the household's expected price level makes workers demand higher nominal wages, so that a higher nominal wage will be set in the wage contracts.[17] It is further assumed that this W^{c} is set for the duration of the labor contract and that workers will supply any amount of labor demanded by firms at W^{c}. That is, for the duration of the wage contract, the *ex ante* labor supply curve is to be temporarily ignored in the analysis and the ostensible labor supply is horizontal, in the neighborhood of the equilibrium, at W^{c} in the (W, n) space.

While the firms negotiate the nominal wage on the basis of their expected price level, profit maximization by the *i* th firm implies, as shown by (29), that its employment and production decision depends only on its own expected price p_i^{ef} rather than on the price level P or the expected price level, so that its employment depends upon W^{c} and p_i. During the production process, the *i* th firm would know the actual price of its own product as a joint element of its production and pricing decision, so that actual employment will be based on the actual prices of the firms' products, rather than on the prices that had previously been expected by the firms. The average of the former is the actual price level, so that the actual aggregate employment n by firms is based on the actual real wage w. In the wage contracts context, this actual real wage is given by the contracted nominal wage W^{c} divided by the actual price level P.[18] That is, the short-run equilibrium level of employments n^* is given by:

$$n^* = n^{\mathrm{d}} = n^{\mathrm{d}}(W^{\mathrm{c}}/P) \quad n^{\mathrm{dJ}} < 0 \qquad (35)$$

Since W^{c} depends upon P^{ef} and P^{eh}, we have:

$$n^* = ©(W^{c}(P^{\mathrm{ef}}, P^{\mathrm{eh}})/P) \qquad (36)$$

where $\partial n^*/\partial P > 0, \partial n^*/\partial P^{\mathrm{ef}} < 0$ and $\partial n^*/\partial P^{\mathrm{eh}} < 0$. The explanation for these signs is as follows. As discussed earlier, increases in the firm's expected price level or/and the household's expected price level establish a higher contractual nominal wage during wage negotiations. This, *ceteris paribus*, – i.e. without an accompanying change in the price level – increases the real wage, which reduces employment. But, for the given contractual nominal wage, an increase in the actual price level lowers the real wage and raises employment. However, © is homogeneous of degree zero in P^{ef}, P^{ef} and P, so that a proportionate increase in the expected and actual price levels will not change employment, even though the nominal wage will rise proportionately.

Employment will thus depend upon the duration of the wage contract, upon the expected price levels by firms and households during wage negotiations, and upon the actual price level when employment occurs.

From (36) and the production function $y=y(n)$, with $y_n > 0$, we have the short-run equilibrium level of employment $y*$ as:

$$y* = \varphi(P^{ef}, P^{eh}, P) \tag{37}$$

where $\partial y*/\partial P > 0, \partial y*/\partial P^{ef} < 0$ and $\partial y*/\partial P^{ef} < 0$. Therefore, for a given contractual nominal wage conditional on expectations,

$$\partial n^*/\partial P > 0,\ \partial y^*/\partial P > 0$$

For most of the commonly used forms of the production and labor supply functions, both $n*$ and y are homogeneous of degree zero in P^{ef}, P^{eh} and P.

Diagrammatic analysis

Figure 14.7a presents the labor demand $n^d(W/P^{ef})$ and the labor supply $n^s(W/P^{eh})$ curves. Note that the vertical axis in this figure is the nominal wage rate W. The negotiated nominal wage will be set at the equilibrium level $W^c{}_0$, and has the expected employment level of n^{*e}_0.

An increase in P^{ef} will shift the labor demand curve to the right and a rise in P^{eh} will shift the labor supply curve to the left, so that each will raise the nominal wage. However, the former will increase the expected employment level and the latter will decrease it. If both P^{ef} and P^{eh} increase proportionately, the two curves will shift proportionately and the nominal wage will increase in the same proportion, without a change in the expected employment level.

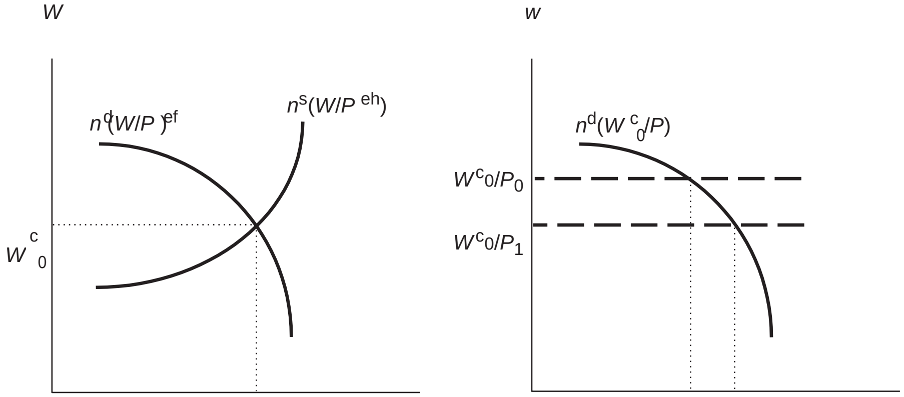


Figure 14.7

Actual employment is determined not in Figure 14.7a but in Figure 14.7b, now with the actual real wage w, equal to W/P, – in the vertical axis. For the contracted nominal wage W_0^c from Figure 14.7a, and a given price level P_0, employment is n_0. With the contracted nominal wage still at W_0^c, a higher price P_1, $P_1 > P_0$, will lower the actual real value of the contracted nominal wage to W_0^c/P_1 and increase employment to n_1 with $n_1 > n_0$. The implicit labor supply curve is horizontal at W_0^c/P in this figure.

If there are no errors in expectations – that is, if $P^{ef} = P^{eh} = P$, actual employment n^* will equal n^{*e}_0, as determined in Figure 14.7a, so that we can take this to be the full-employment level n^f or the "expectational (long-run) equilibrium' level n^{LR}. If P is higher than both P^{ef} and P^{eh}, $n^* > n^{*e}_0$ and vice versa. Therefore, the deviation in employment from its expected level n^{*e} is positively related to the errors $(P - P^e)$ in expectations.

Errors in price expectations, the duration of the wage contract and cost-of-living clauses

This deviation in employment n^* from its expected level n^{*e} will occur only during the duration of the wage contract, since the past errors in expectations will be eliminated when the wage contract is renegotiated. This is usually done through the "catch-up" cost ofliving clauses in labor contracts. Therefore, continuously new errors will be needed to maintain employment above n^{*e}. While this can occur for some time – the "people can be fooled some of the time" syndrome – it cannot continue indefinitely – "they cannot be fooled all the time." The former usually has to take the form of accelerating inflation rates. The latter usually occurs in two ways. One is for the future expectations of inflation to "jump" beyond the past – experienced – inflation rates in an attempt to capture the potential future acceleration in inflation. The other is to reduce or eliminate the loss in purchasing power through inflation by reducing the duration of wage contracts or by building cost-of-living clauses in them. Therefore, while the errors in expectations can induce increases in employment – and do so during accelerating inflation – such increases can only be short term and not a long-term phenomenon in practice. In particular, these increases cannot be relied upon to occur over lengthy periods or persistently high inflation rates.

A simple illustration with linear functions

Assume that the labor demand and supply functions at the time of wage negotiations are:

$$n^s = b_1 W/P^{eh} \qquad b_1 > 0 \tag{38}$$

$$n^d = a_0 - a_1 W/P^{ef} \qquad a_0, a_1 > 0 \tag{38^J}$$

In equilibrium,

$$a_0 - a_1 W/P^{ef} = b_1 W/P^{eh} \tag{39}$$

so that the contractual nominal wage will be:

$$W^c = a_0 \, \Sigma \, \underline{\qquad\qquad} \, \Sigma \qquad P^{ef}$$

$$P^{eh}\ a_1 P^{eh} + b_1 P^{ef} \qquad (40)$$

Hence, $\partial W^c/\partial P^{eh}$, $\partial W^c/\partial P^{ef} > 0$ and a proportionate increase in both expectations increases the nominal wage rate in the same proportion. The expected equilibrium level of employment obtained by substituting this equation in the labor supply function is given by:

$$n^*_e = a_0 b_1 \Sigma \frac{P^{ef}}{a_1 P^{eh} + b_1 P^{ef}} \Sigma \qquad (41)$$

so that a proportionate increase in both expectations does not change n^{*e}. Note that the employment level is not set in the wage contract and can deviate from n^{*e}.

At the time of production, with the nominal wage set by the wage contract at W^c, the actual real wage and employment will be:

$$w^* = a_0 \Sigma \frac{P^{ef} P^{eh}}{a_1 P^{eh} + b_1 P^{ef}} \Sigma \cdot \frac{1}{P} \qquad (42)$$

$$n^* = n^d = a_0 - a_1 a_0 \Sigma \frac{P^{eh} P^{ef}}{a_1 P^{eh} + b_1 P^{ef}} \Sigma \cdot \frac{1}{P} \qquad (43)$$

Both w^* and n^* are homogeneous of degree zero in P, P^{ef} and P^{eh}. If P exceeds both of its expectations, the real wage will turn out to be less than its expectation in the wage contract and n will be greater than n^e – and vice versa. If there are no errors in expectations, $P = P^{ef} = P^{eh}$, so that $n = n^f = a_0$ $a_0 a_1/(a_1$ $b_1)$.+This expectational equilibrium level of employment is independent of the price level and is therefore the classical full-employment level. Note that positive (negative) expectational errors can induce actual employment to be less (greater) than this full-employment level.

The Friedman supply rule

The preceding types of analyses often assume that:

$$P^{ef} = P \qquad (44)$$

This assumption is usually justified by the argument that, for profit maximization, *each* firm only needs to know the price of its own product – in which it possesses a great deal of information – and its own factor costs, represented by the contractual nominal wage, but does not need to know the prices of all the commodities in the economy. Since the average price level P in (42) is only a proxy for the average of the individual commodity prices, each of which is set by the firm supplying the commodity, the firms on average can be expected to predict P fairly accurately.

By comparison, utility optimization by a household requires knowledge of the general price level in order to calculate the purchasing power of the nominal wage. To know the price level requires knowledge of all the commodity prices, which is a degree of knowledge that each household rarely, if ever, possesses. Hence, it is assumed that households individually and on average in the aggregate cannot predict the price level with sufficient accuracy, so that P^{eh} can differ from P.

With these assumptions, the employment and output supply functions can be restated in the more specific form:

$$n^* = \theta(P/P^{eh}) \quad n^{*J} > 0, \partial n^*/\partial P > 0, \partial n^*/\partial P^{eh} < 0 \tag{45}$$

$$y^* = \varphi(P/P^{eh}) \quad y^{*J} > 0, \partial y^*/\partial P > 0, \partial y^*/\partial P^{eh} < 0 \tag{45^J}$$

If $P > P^{eh}$, $w < w^e$, so that labor will prove to be unexpectedly cheaper and firms will employ more than they had expected to employ. Hence, $y^{*J} > 0$, $\partial y^*/\partial P > 0$ and $\partial y^*/\partial P^{eh} < 0$.

(45) and (45^J) are homogeneous of degree zero in P and P^{eh}.

Expectations-augmented employment and output functions

Since the expectations of both firms and households are negatively related to output y, we can simplify the notation by replacing them by the single variable P^e. Therefore, the short-run equilibrium output function becomes:

$$y^* = y(P/P^e) \quad \partial y/\partial P > 0, \partial y/\partial P^e < 0 \tag{46}$$

where P^e is now the expected price level for both firms and households.

There will not be any errors in expectations when $P^e = P$. Designate the respective levels of employment and output when there are no errors in households' price expectations as n^{LR} (or n^f) and y^{LR} (or y^f), consistent with the earlier long-run analysis of this chapter in which there was perfect foresight so that there were no errors in price expectations. These employment and output values are therefore the long-run equilibrium levels of employment and output.

The *log-linear* form of (46) is:

$$\ln y^* = \ln y^{LR} + \beta(\ln P - \ln P^e) \quad \beta > 0 \tag{47}$$

(47) is the *expectations-augmented output function* or *Friedman's supply function*. Note, again, the difference between the superscript LR, which designates the full-employment level (without errors in expectations), and $*$, which designates the short-run equilibrium level in the presence of errors in expectations. Correspondingly, we have for the level of employment:

$$\ln n^* = \ln n^{LR} + \alpha(\ln P - \ln P^e) \quad \alpha > 0 \tag{48}$$

(48) is the *expectations-augmented employment function*.

Note that the price expectations in these equations refer to those incorporated in wage contracts. The economy deviates from its full-employment level due to errors in these expectations; compared with the full-employment level, output is greater if $P > P^e$ and lower if $P < P^e$. In the former case, real wages are lower than the full-employment real wage, making it attractive to hire more labor than in the error-free equilibrium; while the opposite holds in the latter case.

Define *expectational equilibrium* as the state where there are no errors in expectations. That is, it requires that:

$$P^{\mathrm{e}} = P \tag{49}$$

From (47) and (48), in the long-run expectational equilibrium:

$$n = n^{\mathrm{LR}} = n^{\mathrm{f}} \qquad (50)$$

$$y = y^{\mathrm{LR}} = y^{\mathrm{f}} \qquad (51)$$

which assert that in expectational equilibrium the levels of employment and output are the full-employment levels and are independent of the price level. (47) and (48) assert that deviations from these levels occur because of expectational errors, with positive errors ($P > P^{\mathrm{e}}$) causing an increase in output and employment, and negative errors ($P < P^{\mathrm{e}}$) causing a decrease.

The short-run equilibrium unemployment rate and Friedman's expectations-augmented Phillips curve

Unemployment U equals $(L - n)$ and the unemployment rate u equals $U/L = (1 - n/L)$. Therefore, the approximation for the relationship between the log values of u^*, L and n^* is:

$$\ln u^* = \ln L - \ln n^* \qquad (52)$$

From (48) and (52),

$$\ln u^* = \ln u^{\mathrm{n}} - \alpha(\ln P - \ln P^{\mathrm{e}}) \quad \alpha > 0 \qquad (53)$$

where $\ln u^{\mathrm{n}}$ $(= \ln u^{\mathrm{LR}} = \ln L - \ln n^{\mathrm{LR}})$ is the natural rate of unemployment. u^* is the short-run equilibrium unemployment rate in the presence of errors in expectations. (53) is the *expectations-augmented Phillips curve* (EAPC), proposed independently by Friedman (1968) and Phelps (1968) as a correction of the Phillips curve. In a dynamic context, P is replaced by π (the inflation rate) in the Phillips curve equation, so that the usual form of the EAPC is stated as:

$$\ln u^* = \ln u^{\mathrm{n}} - \alpha(\ln \pi - \ln \pi^{\mathrm{e}}) \quad \alpha > 0$$

Note that several assumptions were needed for deriving this equation. Among these was the assumption of market clearance in the labor and commodity markets, an assumption that many Keynesians would not accept. Also note that (53) places the burden of all possible deviations from the natural rate of unemployment on errors in price-level expectations over the duration of the nominal wage contract.

Implications of the expectations-augmented Phillips curve for fluctuations in unemployment

The preceding arguments have the following three implications:

(a) Sources of deviations of the unemployment rate from the natural rate, other than those due to errors in expectations, are ruled out by the EAPC. In particular, the numerous sources of deviations considered by the Keynesians and new Keynesians are not captured.[19]

Among these is the failure of the labor market to clear on a continuous basis, such as in recessions, as well as the possibility of persistent underemployment equilibria in certain rare circumstances such as in the Great Depression of the 1930s. For the new Keynesians, it is the failure to include market imperfections, such as monopolistically competitive firms, that is important (see Chapter 15).

(b) *If* the expectational errors are insignificant in magnitude or are not relevant because of a very short duration of labor contracts,[20] any fluctuations in unemployment over time would have to be explained by a theory of fluctuations in the natural rate of unemployment.

(c) In the EAPC, $u^* < u^{\mathrm{n}}$ requires $w^* < w^{\mathrm{LR}}$ and $P^* > P^{\mathrm{LR}}$, and vice versa, so that fluctuations in employment and output occur because of changes in price level (inflation) but not because of those changes in real aggregate demand which are not reflected in prices (inflation).

Given the experience of considerable cross section and business cycle variations in the unemployment rates in real-world economies, classical theory supplements its theory of expectational errors with a theory of fluctuations in the natural rate of unemployment over time and across countries. In particular, if this theory is to explain the experienced rates of unemployment, it cannot assume that the natural rate is a constant, so that an explanation of the long-run variations in the unemployment rate will require a long-run theory of changes in the natural rate, while an explanation of the cyclical variations in unemployment will require a theory of the cyclical variations in the natural rate.

These arguments lead to two competing sets of theories of unemployment. These are:

1 The classical theories, with fluctuations in the natural rate itself both over the business cycle and the long run, accompanied by continuous labor market clearance but with expectational errors to explain deviations of the actual rate of unemployment from the natural rate. Changes in aggregate demand do not cause fluctuations in unemployment unless they cause prior unanticipated changes in the price level.

2 The Keynesian theories of unemployment, which also allow – but do not require– changes in the natural rate, but emphasize deviations of the actual rate of unemployment from the natural rate due to fluctuations in the demand for commodities and labor, especially in the presence of market imperfections such as contractual rigidities, sticky prices, efficiency wages, etc. Chapter 15 expands on these factors.

Empirical validity of the Friedman supply hypothesis based on errors in price expectations in labor markets

The Friedman output model implies that anticipated monetary policy would not change output and unemployment, which is contradicted by the stylized facts given at the beginning of this chapter. Further, the Friedman model asserts that the effects of monetary policy changes, both anticipated and unanticipated, must go through errors in the price expectations embedded in wage contracts and a subsequent decrease in real wages. This too is contradicted by the stylized facts: an expansionary monetary policy increases output and reduces unemployment without necessarily producing a prior change in the price level or a decrease in real wages. Therefore, Friedman's short-run model of commodity supply does not provide a satisfactory explanation of the short-run impact of monetary policy on output and unemployment.

Price expectations and commodity markets: the Lucas supply function

The EAPC is based on the possibility that unanticipated price changes generate expectational errors in real wages in the labor market. This emphasis on the labor market is in some ways more consistent with the Keynesian aggregative analysis, though the expectations-augmented Phillips curve is associated with Milton Friedman. Neoclassical analysis focuses more on the markets for commodities and its microeconomic Walrasian basis, as in Chapter 3, which implies that the output of each commodity depends upon its relative price. Lucas (1972, 1973) and Sargent and Wallace (1975) modified the certainty version of the microeconomic model by introducing into it uncertainty and firms' expectations on relative product prices. This section presents one version of the Lucas analysis.[21]

Lucas supply function or rule

Assume as in the preceding section that firm *i* produces good X_i, with the quantity produced as x_i and sold at the price p_i. The firm buys inputs, of which labor is taken to be the only variable input. Given the nominal wage W , profit maximization by the firm in perfect competition implies that its supply function is given by:

$$x_i = x_i(p_i/W) \quad x_i^{J} > 0 \tag{54}$$

with p_i determined in the perfectly competitive market *i*. While the firm is not directly concerned with the price level *P*, Lucas's analysis assumes that the nominal wages change proportionately with the price level, so that *P* can be used as an index of labor cost.[22] Therefore, replacing *W* by *P*,

$$x_i = x_i(p_i/P) \quad x_i^{J} > 0 \tag{55}$$

For simplification, Lucas assumes that the price p_i in market i deviates in percentage terms from P by an amount z_i which is normally distributed, independent of P and has a zero expected value and variance η^2. Therefore,

$$p_i = P + z_i \qquad (56)$$

Hence, z_i (p_i P) defines the deviation of the i th product's price from the price level. The product price p_i is called the "local price" while z_i (p_i/P) is referred to as the "relative price,' so that the change in the local price of the i th product incorporates changes in both the price level and its relative price. Both increases in the general price level and in the firms' relative product price can occur at any time. The ith firm is assumed to know the price of its own product but not to know the price level, which it estimates conditional on the information available to it. Given such uncertainty, re-specify (55) as:

$$x_{it} = x_i(p_{it}/E(P^e \,|I\,(i)) \quad x_i^J > 0 \qquad (57)$$

t t

where:

x_{it} = output in market i in period t

e

$E(P_t\, |I_t(i)$ = mean (mathematical expectation) of the price level expected for period t by firms in market i, conditional on $I_t(i)$

$I_t(i)$ = information available in market i in period t.

Specify the log-linear form of (57) as:

$$x_{it} = x^*_{it} + \gamma\,[p_{it} - E(P^e \,|_t I\,(i))] \quad \gamma > 0 \qquad (58)$$

where *all the variables are now in logs* and x^*_{it} is the i th firm's output under perfect certainty or if there are no expectational errors. γ is the firm's response to an increase in its relative price.

Lucas (1972, 1973) provides a specific procedure for determining the expected relative prices. Firms use the available information on aggregate demand and supply movements, and on local and general prices, to form their expectations on the distribution of local and general prices in the present period.[23] This provides them, at the beginning of the current period, with a prior distribution of the expected price level P^e, with mean $\underline{P}$ and constant variance σ^2, with this distribution formed prior to the observation of the current local prices.[24] During period t, the firm knows $\underline{P}$, σ^2, η^2 and observes p_i. The (ith) firm uses this knowledge of its local price to calculate $E(P^e_t|I_t(i))$ as:

$$E(P^e \,|I\,(i)) = \underline{P}_t + [\sigma^2/(\eta^2 + \sigma^2)][p \;\; - \underline{P}_t] \qquad (59)$$

t t it

where:

$\underline{P}_t$ = mean of the prior distribution of expected prices

σ^2 = expected variance of P ("price level variability")

η^2 = expected variance of z_i ("relative price variability")

$\sigma^2 + \eta^2$ = expected variance of p_i ("local price variability").

The second term on the right of (59) is the correction made to the prior expectation $\underline{P}_t$ of the price level on the basis of the observed local price p_{it} .

Equation (59) was justified on the basis of information available in the *i* th market when the price level is not directly observed. This view of the nature of information was couched by Lucas (1972) in what has come to be known as the *island parable*. This parable envisions the workers and firms as distributed spatially over islands (or isolated points). The firms do not know about activity (prices and output) on other islands but must forecast the average price level (over all the islands) in order to formulate their labor demand and output supply decisions. To forecast the price level, they use the historic variability of their island price – represented by $(\eta^2 + \sigma^2)$ – relative to overall variability to forecast the shift of the price level from a prior expected level, and do so as specified in (59).

Rewrite (59) as:

$$E(P^{e}_{t} \,|\, I_{t}(i)) = \alpha\underline{P}_t + (1-\alpha)p_{it} \qquad (60)$$

where $\alpha = \eta^2/(\eta^2 + \sigma^2) \geq 0$, so that α is the expected ratio of the relative price variance to the total local price variance.

Substituting (60) in (58), we have:

$$x_{it} = x^*_{it} + \alpha\gamma\,[p_{it} - \underline{P}_t] \qquad (61)$$

Integrating (61) over all markets *i*, with total output supply designated as y^s, and replacing the known local price p_{it} by the *actual* aggregate price level P_t ,

$$y^s_t = y^*_t + \alpha\gamma\,[P_t - \underline{P}_t] \qquad (62)$$

which is the *aggregate supply function* based on firms' expectations about variations in local and general prices. The two components of a firm's response to an expectation error are: (a) γ, which is the change in output in response to the expectation error (P P^e); and (b) α, which is the revision of P^e from its prior value $\underline{P}$. (62) is known as the *Lucas ~~-~~supply function or rule*.

If $\eta = 0$ and $\alpha = 0$, so that relative prices are expected to be stable, (62) becomes:

$$y^s_t = y^f_t \qquad (63)$$

In this case with $\eta = 0$, aggregate supply will not respond to absolute price changes and hence to changes in aggregate demand, with the result that the aggregate supply function will be vertical in the (y, P) space. But if $\sigma = 0$ – that is, the price level is expected to be stable, so that the change in local price is taken to be wholly a relative price change – the aggregate supply function will become:

$$y^s_t = y^*_t + \gamma\,[P_t - \underline{P}_t]$$

where γ is likely to be positive since it reflects firms' responses to an increase in their relative prices.

Another explanation of the Lucas supply function

Equation (62) was based on price misperceptions. It is to be distinguished from a somewhat similar equation, also attributed to Lucas, which can be derived from the intertemporal substitution of work in household decisions. In this model of intertemporal utility maximization, for given nominal wages in each period and a given current price level P_t , if the expected price level (P^e_{t+1}) rises, it will decrease the expected future real wage (w^e_{t+1}) for the given nominal wage,[25] while the current real wage (w_t) is unaffected. Hence, from utility optimization, workers will substitute work in the current period (i.e. increase n^s_t by decreasing their leisure time in t) for work in future periods (i.e. decrease n^s_{t+} by increasing leisure in $t+1$). Conversely, they would work – and produce – less in the present period if, at the given current price level, the expected future price is lower and the expected real wage higher. Such behavior would produce recessions in output in the latter case and booms in the former, thus causing real business cycles. However, the empirical significance of such a model is limited since the observed intertemporal substitution of work in labor supply decisions on the basis of price movements is too low to imply the larger observed variations in output over the business cycle, so that we shall ignore it further in this chapter.

Comparing the Friedman and the Lucas supply functions

We are, therefore, left with two types of expectations-based Phillips curve relationships. One of these is Friedman's expectations-augmented Phillips curve, which is based on errors in expectations in labor market and contractual rigidities; and the other is Lucas's supply function, based on errors in the expectations of relative prices in commodity markets. The former is sometimes claimed to be an example of the latter under the argument that nominal wages are the "local" price of workers as suppliers of labor and the observed increases in these nominal wages are used by the workers in forming their expectations on the price level and the real wage (the "relative price" of labor). However, the theoretical and empirical bases of the two are quite different,[26] so that it is preferable to keep them separate for analytical reasons. But they are similar in spirit in that both maintain full employment unless there are errors in expectations. Note that the errors in price expectations will be corrected as time passes since this will provide information on the actual prices. In modern economies, the lag in information on actual prices and inflation rates is often a month or a few months, so that any expectational errors would be corrected fairly soon. Therefore, both the errors in expectations and the deviation of short-run equilibrium output from the full-employment output will be *self-correcting* and *transient.*

The Lucas model with supply and demand functions

The Lucas supply function derived above is:

$$y^s_t = y^*_t + \alpha\gamma\,[P_t - \underline{P}_t] \tag{62}$$

On the *aggregate demand function*, Lucas (1972, 1973) assumed that:

$$Y^d_t = Y^d_{t-} + \delta + \mu_t \tag{64}$$

where:

Y^{d} = nominal aggregate expenditures/demand
δ = systematic (known) increase in demand
μ = random shift in demand, with $E\mu = 0$

Further, from the definition of nominal expenditures and noting that all variables are in logs:

$$Y_t = y_t + P_t \tag{65}$$

Assuming equilibrium in the commodity market with $y^{d} \quad y^{s} \quad y$ and $Y^{d} \quad Y^{s} \quad Y$, eliminate y_t from (62) to (65). Then, assuming rational expectations as applied by the classical paradigm (which is that the expected level of a variable is its long-run equilibrium level), $Ey_t = y^{LR}_t$ and $\underline{P} = EP_t$, we get:

$$P_t = \frac{\alpha\gamma\,\delta}{1+\alpha\gamma} + \frac{1}{1+\alpha\gamma}Y_t + \frac{\alpha\gamma}{1+\alpha\gamma} Y_{t-1} - y^{LR}_t \qquad \alpha \geq 0,\ \gamma > 0 \tag{66}$$

Starting from (65), substitute (64) for Y_t, and (66) for P_t.. This yields:

$$y^*_t = y^{LR}_t - \frac{\alpha\gamma\,\delta}{1+\alpha\gamma} + \frac{\alpha\gamma}{1+\alpha\gamma}(Y_t - Y_{t-1}) \quad \alpha \geq 0,\ \gamma > 0 \tag{67}$$

Given (64), (67) in equilibrium (with $y^{d} = y^{s} = y$) simplifies to:

$$y^*_t = y^{LR}_t + \frac{\alpha\gamma}{1+\alpha\gamma}\mu_t \quad \alpha \geq 0,\ \gamma > 0 \tag{68}$$

This equation forms the theory of output of the modern classical school. It shows that the modern classical school does not assume full employment or maintain its continuous existence. It does assert that deviations in output from the full-employment level can be caused by errors in price expectations; however, any such deviations will be self-correcting and transient under rational expectations. The causes of the deviations of actual output from short-run equilibrium are ruled out by the way the analysis is formulated. In this theory, the only fluctuations in output that are not transitory and self-correcting have to come from fluctuations in full-employment output due to shifts in technology and physical capital, labor supply and human capital, and availability of resources. Real business cycle theory rests on the preceding Lucas model and expands on this theme.

If firms believe that all the variation in prices is in the general price level and none in the relative price level, $\eta = 0$, so that $\alpha = 0$ and (68) yields the error-free output as $y^{LR}t$. That is:

$$y^*_t = y^{LR}_t \quad \text{for } \eta = 0$$

so that $\partial y_t / \partial\delta = \partial y_t / \partial\mu = 0$. Therefore, (68) implies that, if firms do not perceive any changes in relative prices, both systematic and random shifts in aggregate demand will not cause deviations in output from its full-employment level under certainty. But if relative prices are expected to change, $\eta >$ 0 and >0. In this case, random – but not systematic – shifts in aggregate demand can have real effects, since from (68),

$$\partial y^*_t / \partial \delta = 0 \quad \text{for} \quad \eta > 0 \tag{69}$$

while:

$$\frac{\partial y^*_t}{\partial \mu_t} = \frac{\alpha\gamma}{1+\alpha\gamma} = \frac{\eta^2\gamma}{\sigma^2+\eta^2(1+\gamma)} > 0 \text{ for } \eta > 0 \qquad (70)$$

Note from (70) that even the *random* shifts in aggregate demand cause changes in output *only* if firms misinterpret its impact on the price level and believe that some part of the resulting increase in prices is a relative price increase. In conditions of hyperinflation where the likelihood and magnitude of general price increases dominate over relative price increases, the public's expectation is likely to be $\eta = 0$, so that even random changes in aggregate demand will not have any effects on output. In this context, neither systematic nor random demand increases will change output. Therefore, for hyperinflations, money is likely to be neutral for both systematic and random increases in the money supply. Hence, even random increases in the money supply will not always induce increases in output and employment.

Asymmetric information and the impact of systematic demand increases on output

Equation (69) states that systematic demand increases by the policy makers will not change real output. But, suppose that the policy makers systematically (for example, through the use of a rule such as the Taylor rule) increase demand but in such a way (as by a one-time shift in the coefficients of the rule followed) that the firms interpret it to be a random increase. This is a case of asymmetric information between the policy maker and the public. This "disguised or misinterpreted" nature of systematic demand increase will allow the policy maker to increase real output through the "random multiplier" $\partial y/\partial \mu$, whose value is specified by (70). However, the systematic nature of the demand increase will, sooner or later, be observed by firms, leading to two types of change. One, firms will find it optimal to acquire better information so that their likelihood of correctly perceiving a systematic demand increase as being systematic – rather than erroneously as random – will increase. For this correctly perceived systematic demand increase, there will be no change in output. Second, firms faced with repeated increases in the general price level will modify their expectations on prices by increasing the value of σ^2 relative to η^2, so that $\partial y/\partial \mu$ will decrease. In the limiting case, if there were only systematically induced increases in the price level, firms would adjust their expectations such that $\eta^2/(\sigma^2 \ \eta^2) \ 0$, so that $\partial y/\partial \mu$ will go to zero. That is, systematic demand increases, disguised or misinterpreted as being random ones, will eventually lose their efficacy.

The implications of the Lucas model for unemployment

To examine the response of unemployment to changes in aggregate demand, start with the definition of unemployment as being ($L \quad n$) and use the production function $y \quad y(n)$, $y^J(n) > 0$, to go from y to u. Then, (68) implies that employment n is a function of the random demand term μ_t but not of the systematic demand term δ. With all variables being in logs, and assuming a log-linear relationship, (68) implies that:

$$u^*_t = u^n - \beta\mu_t \quad \beta \geq 0 \qquad (71)$$

where u^* is the short-run equilibrium unemployment rate, β is a positive function of

$[\alpha\gamma/(1+\alpha\gamma)]$ with $\beta = 0$ if $\alpha = 0$. First note that δ is not in (71), just as it was not in

(68), so that $\partial u_t^*/\partial\delta = 0$. Second, note that α is in (71) through β, just as it was in (68), and that $\beta > 0$ for $\alpha > 0$ and $\gamma > 0$. For $\beta > 0$, unemployment decreases if there is a random increase in aggregate demand, since the latter would persuade individual firms that the relative demand and the relative price of their product have increased. That is, for $\beta > 0$, (71) implies a negatively sloped curve in the (y, u) space, which looks like a Phillips curve – but is not the Phillips curve[27] – and can be called the "Lucas–Phillips curve." Hence, there is a seeming tradeoff – for *given* expectations on general versus relative price increases – between price level increases and output increases in the short run.

However, there is a vital difference between (71) and the standard Phillips curve (for which, see Chapter 15). While the latter was envisaged by the Keynesians as a durable tradeoff between price increases (both anticipated and unanticipated) and unemployment, (71) cannot be used as a durable tradeoff for policy purposes. Meaningful policy increases in demand cannot be random but must be systematic, such as by a constant δ in (64) above or according to some established rule, and, under the Lucas analysis, any systematic demand changes cannot change real output. Further, as argued above, any systematic demand or price increases, disguised or misinterpreted as being random ones, will sooner or later lose their efficacy as their systematic nature becomes understood.

Hence, according to (71), correctly anticipated inflation – such as would be the case if the inflation rate were constant or steadily increasing – has a vertical Lucas–Phillips curve, thereby not providing a durable impact of money, prices and inflation on output. This is now generally accepted. However, it is now also accepted that in the short run, monetary policy, whether operating through money supply or interest rates, does affect output and does so earlier than prices and inflation, as indicated in the stylized facts at the beginning of this chapter. It is now generally accepted that these facts cannot be explained in a satisfactory manner by imperfect information, resulting in temporary price or inflation misperceptions, but must rest on theoretical foundations and empirical factors not encompassed in Lucas's analysis.

Empirical validity of the Lucas supply model based on price misperceptions in the commodity market

The Lucas output model (Lucas, 1972, 1973; Sargent and Wallace, 1976; see also Chapter 17) implies that anticipated monetary policy would not change output and unemployment. This clearly contradicts the stylized facts given at the beginning of this chapter. Further, the Lucas model asserts that the effects of monetary policy changes, both anticipated and unanticipated, must go through price level changes and errors in price expectations. This too is contradicted by the stylized facts: an expansionary monetary policy increases output and reduces unemployment without necessarily producing a prior change in the price level, as Lucas (1996) concluded later from his assessment of the empirical evidence (see Section 14.16). Therefore, for the short run, the Lucas model does not provide a satisfactory explanation of the impact of monetary policy on output and unemployment.

Defining and demarcating the models of the classical paradigm

Evolution of the classical paradigm

The principles of the classical paradigm evolved out of several rather disparate elements. The dominant one was the economic and political philosophy of liberalism in the first half of the nineteenth century. Its economic analysis was that of individual markets for commodities, factors of production, money and bonds, with competition as the invisible hand guiding each of them to equilibrium through the adjustment of the relevant price. In this analysis, all prices were flexible and adjusted to bring each market into equilibrium, so that each market cleared at the trading price. Much of this analysis of individual markets was set out in the following rather separate categories: individual commodity and labor markets to determine relative prices, wages, output and employment; quantity theory for determining the price level and the loanable funds theory for determining the interest rate; and business cycle theories. These distinctive theories nowhere appeared in an integrated format so that one cannot point to the work of any economist prior to Keynes's book *The General Theory* (1936) for a statement of an integrated macroeconomic theory or model. In particular, there was no integrated model that included the consumption and saving functions and the investment multiplier, and therefore, no aggregative theory of the commodity market. Further, while there was a great deal of discussion about the nature of risk and uncertainty, the above components of the traditional classical approach did not incorporate it in a meaningful way. These are the elements of what Keynes labeled in Chapter 1 of *The General Theory* as the classical model, though there did not exist in the literature at that time such an established and complete macroeconomic model. We have labeled this pre-Keynesian model the *traditional classical model*, in comparison with the subsequent neoclassical and modern classical models.

Keynes's *The General Theory* for the first time provided an integrated macro model of the economy and also fundamentally altered the way the profession models short-run aggregative economics. He chose to give the foremost place in his model to the commodity market, with an analysis incorporating the multiplier and also money demand analysis. John Hicks (1937) organized Keynes's ideas on the commodity and monetary sectors into what he labeled the IS–LM framework, and cast the traditional classical model in the same format to facilitate comparison. The traditional classical model thus recast in the mould of the IS–LM framework, which had been proposed to illuminate Keynes's ideas, came to be known as the neoclassical model. The IS–LM model of aggregate demand, combined with the AD–AS model for the determination of output and the price level, provided the first formally integrated macroeconomic model of the classical paradigm. It was elaborated and refined in the decades up to 1970.

Neoclassical economics with the addition of rational expectations, if there is uncertainty, and an insistence on continuous labor market clearance has been labeled in this book the *modern classical model*.[28] The combination of the modern classical model with Ricardian equivalence has been labeled the *new classical model*. In general, these models of the classical paradigm do not incorporate market imperfections, that is, deviations from the perfect markets assumption, which are emphasized in Keynesian economics (see next chapter).

There can be considerable disputes about the proper delineation of the classical schools or models. We introduced the following taxonomy in Chapter 1. We elaborate on it here, though at the risk of some repetition. No claim is being made to this taxonomy being a universal – or perhaps even a majority – one. We have chosen it below for reasons of clarity in separating each model from the others, rather than leaving their differences ambiguous, while maintaining consistency with the writings and folklore in the history of economics thought.

All the schools of the classical paradigm share the common belief that the real-world economy under consideration – and not just the models – functions at full employment in the long run and that one of the characteristics of long-run equilibrium is the independence of the real variables from the financial ones, so that money is neutral in the such equilibrium. Further, all schools share the belief that deviations from the long-run equilibrium can occur in the short run but such deviations are self-correcting and transient. States with less than full employment are, therefore, states of disequilibrium during which the economy continues to adjust towards its full-employment equilibrium – and not away from it, A major difference among these schools is whether the real-world economy adjusts so fast as to have continuous equilibrium, so that it will not show any evidence of disequilibrium, even though disequilibrium remains a hypothetical state within the model.

The traditional (pre-Keynesian) classical ideas

This section lays out our interpretation and distillation of the writings of the pre-Keynesian economists. Their ideas were not expressed or formulated in terms of a compact model, and the analysis of the expenditure sector (the IS curve) and the multiplier was not available to them. Their common belief was that output and employment in the long run equilibrium depended upon the real sector's relationships only and were independent of the monetary sector. The economy's interest rate was determined by the theory of loanable funds, which in modern terminology corresponds to the market for bonds (see Chapter 19). Further, the quantity theory for the determination of the price level applied in equilibrium. But outside this equilibrium, changes in the money supply could change output and employment. Not only could the money supply affect the real sector in this way, the economy was considered to be very prone to fluctuations in output and employment. Many of these fluctuations were attributed to money supply shocks or the response of the money supply to real shocks. In particular, most of the classical economists did not believe that the economy functioned so well that it always maintained full employment or that it did so most of the time. In fact, recessions and crises – many of them originating in the banking sector or financial speculation or occurring due to the response pattern of the financial sector to real shocks – were common, and widely recognized as such, during the nineteenth century. Hence, the traditional classical school did not assume continuous full employment or that it existed most of the time.

The neoclassical model

This model was an attempt to bottle the main ideas of the pre-Keynesian classical economists into a compact modern macroeconomic model. This process was initiated by Hicks in 1937, who borrowed the IS–LM analysis from Keynes's work and used it as a technique for interpreting the traditional classical ideas, thereby re-incarnating those ideas into the neoclassical model. The neoclassical model continued to have both equilibrium and disequilibrium aspects and did not assume instantaneous market clearance. In this, it represents the ideas of the pre-Keynesian classical economists more faithfully than does the modern classical model.[29]

The modern classical and new classical models

The certainty version of the modern classical model modifies the neoclassical model by adding the assumption of continuous market clearance, especially of the labor market at the full-employment level. By doing so, it ignores the disequilibrium properties and multipliers of the neoclassical model as being irrelevant in practice. The uncertainty version of this model adds in the rational expectations hypothesis. Some economists would also add to this mix the assumption of Ricardian equivalence. However, our definition of the modern classical model excludes this assumption, only making it part of the new classical model. This differentiation means that, under our definitions, fiscal policy would change aggregate demand in the modern classical model but not in the new classical model.

Hence, under our designations, the constituents of the *modern classical model* are:

1 the neoclassical model, modified by the additions of:
2 uncertainty, with deviations of output and employment from their long-run values due to errors in price expectations;
3 the rational expectations hypothesis, which implies that the errors in expectations will be random;
4 continuous market clearance (especially of the commodity and labor markets).

The constituents of the *new classical model* are:

1 the modern *classical* model, modified by the addition of:
2 Ricardian equivalence.

Note that both the modern classical and the new classical models do possess money supply changes as a policy tool for changing aggregate demand in the economy, but, of the two, only the modern classical model allows fiscal policy to change aggregate demand. For the long run, both these models imply the neutrality of money in the full-employment state, so that the impact of the money supply and velocity changes can only be on the price level and not on real output and employment. Further, for the short run, unanticipated changes in money supply and velocity can cause deviations of short-run output and employment from their long-run values, but, given rational expectations, these deviations will be transient and self-correcting as new information on prices becomes available. Therefore, both the modern classical and the new classical models imply that there is neither a need nor scope for systematic monetary policy for changing the levels of output and employment in the economy, so that such policies should not be pursued. These ideas are further explored in Chapters 15 and 17.

Real business cycle theory and monetary policy

Business cycles are cyclical fluctuations in the economy's output and employment in real, not analytical, time. Their explanation relates to the short term, which is a chronological concept of time, rather than the analytical short run or long run.

Real business cycle theory is an offshoot of the modern classical model and asserts that business fluctuations occur *only* in response to shocks to the fundamental determinants of long-run output and employment (e.g. see Prescott, 1986; Christiano and Eichenbaum, 1992; Romer, 1996, Ch. 4). These determinants are technology, which determines the production function and the demand for inputs, and the supply of factor inputs. Among the determinants of the latter are preferences, including those on labor supply, which depends on the labor–leisure choice and the stock of resources. Shifts in the production function or input supplies alter long-run equilibrium output, as well as being a source of cyclical fluctuations in output. The real business cycle theory derives the fundamental determinants of business cycles from the general macroeconomic models of the classical paradigm.

Explicitly, or by omission, real business cycle theory also holds that shifts in aggregate demand, no matter what their source, do not cause changes in output and employment and therefore do not cause business cycle fluctuations. Therefore, changes in consumption, investment, exports, money supply and demand (or the central bank's interest rate policy) or fiscal deficits cannot change output and employment. This exclusionary proposition is derived from the properties of the long-run equilibrium of the modern classical model. To be valid, it requires perfectly competitive markets and also that long-run equilibrium is *continuously* maintained in the economy.

The policy implication of real business cycle theory, as of the modern classical model of which it is an elaboration, is that systematic monetary (and fiscal) policies cannot affect output and employment, so that they cannot be used to moderate the business cycle. The critical elements for this implication are the Friedman–Lucas supply equation and rational expectations, according to which anticipated changes in prices, inflation and monetary policy cannot affect output. Therefore, the Taylor rule, under which systematic monetary policy manipulates aggregate demand by changing the interest rate in response to the output gap and the deviation of inflation from its target rate, can only be useful in controlling inflation but not in moderating the output gap. According to the modern classical school, while random monetary policy can change aggregate demand, the central bank cannot predict and therefore cannot offset the random fluctuations in the private components of aggregate demand. In short, in the new classical model, monetary policy and the Taylor rule have no legitimate role in moderating or reducing the duration of business cycles.

Intuitively, the problem with the real business cycle theory is most evident in its explanation of recessions. It attributes recessions to a fall in labor productivity and/or an increase in the preference for leisure. The objections to these explanations are succinctly stated by the quip: recessions occur because "workers forget how to do things" ("lose some of their knowledge") and/or because they decide to become lazier for some time, thereby causing the recessionary fall in output! Neither of these explanations is plausible, so the validity of the real business cycle theory is highly doubtful. Looking at upturns in business cycles, the real business cycle theory attributes upturns to increases in productivity and/or increases in the preference for work over leisure. The latter is hardly plausible over the length of upturns in the economy, while the former is highly plausible. Here, however, it is the plausibility of the assertion of real business cycle theory that aggregate demand increases cannot also be a source of upturns that is highly doubtful.

The real business cycle propositions rest on the assumption that all markets can be taken to be competitive and efficient (i.e. continuous equilibrium) in the economy. This assumption is not consistent with models of the Keynesian paradigm, since they incorporate market imperfections and/or failure of the economy to achieve long-run equilibrium instantly after a demand shock. In these models, shifts in aggregate demand, whether through shifts in investment and other private sector variables or in monetary and fiscal policy, can produce changes in output and be a source of, or contribute to the continuation of, business cycles. More specifically on monetary policy, market imperfections can create non-neutrality of money, so that fluctuations in the money supply can add to output fluctuations. Conversely, the appropriate monetary policy can reduce the severity of cyclical fluctuations due to aggregate demand shocks coming from the private sector. Further, Keynesians do not deny that shifts in the fundamental determinants of output, mentioned above, can also cause output fluctuations.

Therefore, the core of the debate about the validity of real business cycle theory is not about whether shocks to technology and factor inputs can cause cyclical fluctuations, for that is not in dispute. It is rather about whether shocks to aggregate demand can cause such fluctuations and whether monetary policy can moderate them. Real business cycles and the modern classical school deny that they can, or do so in a significant manner, while Keynesians assert that they can do so. This issue is easily testable by the appropriate causality tests. The consensus on the empirical evidence seems to be that the major part (in some estimates, as large as 70 percent) of the fluctuations in output can be attributed to productivity shocks. This is a testament to the success of real business cycle theory, as compared with Keynesian ideas from the 1940s to the 1970s that had attributed most business cycle fluctuations to shifts in aggregate demand. However, the empirical evidence leaves a very significant part of the fluctuations in output that cannot be explained by shifts in technology and preferences. Overall, the empirical evidence, as well as intuition, seems to indicate that fluctuations in aggregate demand, in addition to changes in technology and preferences, do cause fluctuations in output and employment and that money supply growth is positively related to output growth. Therefore, real business cycle theory is not strictly valid, and monetary policy can be pursued in appropriate cases to reduce output fluctuations.

The exponents of the real business cycle theory also prefer to test this theory by the calibration and simulation of models rather than by the econometric testing of their hypotheses. The former procedure requires a priori specification of the likely values of the parameters, on which there can be considerable doubts. Further, the findings may not be robust to small changes in these assumed values, or consistency with the empirical observations may require implausible values. Consequently, this testing procedure and its reported findings have not won general acceptance.

There seem to be at least two major contributions of the real business cycle theory. One, it has firmly established that changes in technology and preferences do cause cyclical fluctuations in output and may do so significantly more than fluctuations in aggregate demand. Two, the approach initiated by the real business cycle agenda to macroeconomic modeling is now firmly established. This approach requires that macroeconomics be based on optimization over time by individual economic agents in a dynamic context. This stochastic dynamic intertemporal approach to macroeconomics permeates current macroeconomic models, including the new Keynesian model, which is presented in the next chapter. The major deficiency and unrealistic assertion of the real business cycle theory is that it denies demand shifts any role in output fluctuations.

The empirical evidence on the impact of changes in aggregate demand on output is often on the impact of money supply changes, which change aggregate demand, on output. The influential study by Friedman and Schwartz (1963a,b) used evidence from over 100 years of US data to show clear evidence that money supply changes lead, and therefore Granger-cause, changes in real economic activity. However, inside money (i.e. deposits in banks) is the largest component of money. Subsequent contributions by other authors showed that deposits respond to macroeconomic disturbances, so that money is more highly correlated with lagged output than with future output; i.e. deposits lag rather than lead output. However, monetary aggregates such as M2 still lead output. Further, if the central bank uses the interest rate as its operating monetary policy target, and money supply responds endogenously to it, the evidence seems to show that changes in interest rates lead output.

To conclude, empirical evidence shows that while shocks to real factors such as technology and preferences do cause fluctuations in output, shocks to monetary policy variables of money and interest rates also do so. Models of the modern classical school and real business cycle theory do not provide a satisfactory explanation for the latter finding. In recent years, sticky price and inflation models of the new Keynesian school have been proposed to explain economic fluctuations. An example of these studies is provided by Ireland (2001b).

Milton Friedman and monetarism

Milton Friedman occupies a special place in the counter-reformation from Keynesian economics to the neoclassical and eventually to the modern classical theories, though his ideas are, in many ways, closer to the neoclassical economics of the 1960s and 1970s than to the modern thinking. In the 1950s, Friedman argued and showed through his theoretical and empirical contributions that “money matters” – that is, changes in the money supply change both nominal and real output – is against the then general view of the Keynesians that changes in the money supply brought about through monetary policy did not significantly affect the economy, or did so unpredictably.[30] He argued and tried to establish through empirical studies that, as far as nominal national income was concerned, the money-income multiplier was more stable than the investment-income multiplier, so that monetary policy was more predictable than fiscal policy in its impact on nominal national income. However, Friedman held that major instability in the US economy had been produced or, at the very least, greatly intensified by monetary instability and that major depressions were associated with monetary contractions, prior to and after the establishment of the Federal Reserve System in 1913. Therefore:

> The first and most important lesson that history teaches about what monetary policy can do … is that monetary policy can prevent itself from being a major source of economic disturbance. [However, while] monetary policy can contribute to offsetting major disturbances in the economic system arising from other sources … we simply

> do not know enough to be able to recognize minor disturbances when they occur or to be able to predict what their effects will be with any precision … [so that monetary policy should only offset major] disturbances when they offer "a clear and present danger.'
>
> (Friedman, 1968, pp. 12–14).

Friedman is famous for his assertion that inflation is always and everywhere a monetary phenomenon and that increases in the money supply will produce inflation, not increases in real output, in the long run. Another aspect of Friedman's agenda to re-establish the doctrine that money "matters" for short-run fluctuations in output and employment was to set out in the 1950s and 1960s the theory – and to establish empirically – that money demand was a function of a few variables and that the money demand function was stable, with the result that the velocity of money also had a stable function. We have already discussed some of these contributions in Chapter 2 in the context of Friedman's restatement of the quantity theory of money. These arguments had been accepted by the profession by the early 1960s, and contributed to the conversion of Keynesian macroeconomics to a Keynesian–neoclassical synthesis expressed by the IS–LM model for the determination of aggregate demand.

On the relationship between the nominal variables and the real side of the economy, Keynesians in the late 1950s and 1960s had relied on the Phillips curve, which showed a negative tradeoff between the rate of inflation and the rate of unemployment. Friedman argued that the natural rate of unemployment – and, therefore, full-employment output – was independent of the anticipated rate of inflation, so that the fluctuations in output and the rate of unemployment were related to deviations in the inflation rate from its anticipated level. This relationship came to be known as Friedman's expectations-augmented Phillips curve and incorporated his contributions on the natural rate of unemployment.

While Friedman brought the role of anticipations on the rate of inflation into discussions on the role and effectiveness of monetary policy in the economy, he did not use the theory of rational expectations; the rational expectations hypothesis had not yet entered the literature and Friedman relied on adaptive expectations in his empirical studies.

Hence, Friedman was a precursor of the modern classical school but not fully a member of it. Nor does this school follow all of his ideas. He was closer to the Keynesians in one important respect than to the later modern classical school. He believed, as did the Keynesians, that the economy does not always maintain full employment and full-employment output – and does not always function at the natural rate of unemployment, even though this concept was central to his analysis. Hence, policy-induced changes in aggregate demand could induce short-term changes in output and employment. Therefore, money mattered even to the extent that changes in it could induce changes in employment and output, depending upon the particular stage of the business cycle. While this view was shared with the Keynesians, Friedman tilted against the Keynesians on the pursuit of discretionary monetary policy as a stabilization tool – especially for "fine tuning" the economy – because of his belief that the impact of money supply changes on nominal income had a *long and variable* lag. He reported on the lag in the impact of monetary policy that:

> The rate of change of the money supply shows well-marked cycles that match closely those in economic activity in general and precede the latter by a long interval. On the average, the rate of change of the money supply has reached its peak nearly 16 months before the peak in general business and has reached its trough over … 12 months before the trough in general business. … Moreover, the timing varies considerably from cycle to cycle – since 1907 the shortest time span by which the money peak preceded the

> business peak was 13 months, the longest 24 months; the corresponding range at troughs is 5 months to 21 months.
>
> (Friedman, 1958; see Friedman, 1969, p. 180).

With such a long and variable lag between changes in the money supply and nominal income, the monetary authorities cannot be sure when a policy-induced increase in the money supply would have its impact on the economy. Such an increase in a recession may not, in fact, increase aggregate demand until the following boom, thereby only increasing the rate of inflation at that time. Consequently, Friedman argued that discretionary monetary policy, intended to stabilize the economy, could turn out to be destabilizing. Friedman's recommendation on monetary policy was, therefore, that it should maintain a low constant rate of growth, as stated by him in:

> There is little to be said in theory for the rule that the money supply should grow at a constant rate. The case for it is entirely that it would work in practice. There are persuasive theoretical grounds for desiring to vary the rate of growth to offset other factors. The difficulty is that, in practice, we do not know when to do so and by how much. In practice, therefore, deviations from the simple rule have been destabilizing rather than the reverse.
>
> (Friedman, 1959).[31]

Therefore, while both Friedman and the modern classical economists are opposed to the pursuit of discretionary monetary policy, they arrive at this position for quite different reasons. For Friedman, money supply changes can change output and employment but the long and variable lags in this impact make a discretionary policy inadvisable; over time; it could make the economy perform worse rather than better. For the modern classical economists, the economy maintains full employment except for transitory and self-correcting deviations from it due to random errors in price expectations, so that systematic policy changes in the money supply cannot change output and employment, but only the price level. Further, for this school, the lags in the impact of systematic money supply changes on nominal national income are not significant.

On the transmission mechanism from money supply changes to income changes, Friedman supported Fisher's direct transmission mechanism (from money supply changes directly to expenditures) over the indirect one (from money supply to interest rates to investment in the Keynesian and IS–LM models). Neoclassical and modern classical models espouse the latter rather than the former.

Monetarists and the St Louis equation

Monetarism and monetarists have been defined in a variety of ways. In a very broad sense, monetarism is the proposition that "money matters" in the economy. In this sense, Friedman, Keynes and the Keynesians[32] were all monetarists, while the modern classical

[29] The above prescriptions for policy were repeated in Friedman (1968).

[30] This was not true of many Keynesian models popular in the 1950s and 1960s. Some of these relegated money to a minor role, since they claimed that money was only a small part of the economy's liquidity, which included trade credit, etc. Other models claimed that the price level and the inflation rate were an outcome of the struggle between

school is less monetarist since it downplays the impact of money supply changes on the real variables of the economy. In a narrow sense, monetarism as a label was associated with the St Louis school in monetary and macroeconomics. We shall define monetarism in this narrow sense. The St Louis school provided in the late 1960s and early 1970s an empirical procedure for estimating the relationship between nominal income and the money supply. This was the estimation of a reduced-form equation (Andersen and Jordan, 1968) of the form:

$$Y_t = \alpha_0 + \Sigma_i a_i M_{t-i} + \Sigma_j b_j G_{t-j} + \Sigma_s c_s Z_{t-s} + \mu_t \qquad (72)$$

where:

Y = nominal national income
M = nominal value of the appropriate monetary aggregate
G = value of the appropriate fiscal variables
Z = vector of the other independent variables
μ = disturbance term.

Equation (72) is called the St Louis monetarist equation and was presented earlier in Chapter 8. While its common form used nominal income as the dependent variable, the dependent variable can be changed, depending upon the researcher's interest, to real output, the unemployment rate, the rate of inflation or some other endogenous variable. In general, the St Louis equation is a reduced-form estimation equation of the short-run macro models, with the monetary aggregates and the fiscal variables being taken as exogenous.

The St Louis equation has become a popular method for determining the impact of monetary and fiscal policies on nominal national income and other variables. Its initial estimation by researchers at the Federal Reserve Bank of St Louis (Andersen and Jordan, 1968) showed that the money aggregates had a strong, positive and rapid impact on nominal income, this impact being more significant than that of fiscal policy. The marginal money-income multiplier was about 5 over five quarters, while the marginal impact of fiscal policy was positive for the first year and then turned negative, with a multiplier of only about 0.05 over five quarters.[33] These findings were consistent with Friedman's stance, except that the estimations of the St Louis equation indicated a much shorter and more reliable lag than Friedman had found. Therefore, contrary to Friedman's recommendations and consistent with those of the Keynesians, monetarism was consistent with the stance that monetary policy could be useful for short-term stabilization.

The St Louis monetarism represented a transitional stage in the transition from Keynesian ascendancy in economics in the decades before 1970 to the ascendancy of the neoclassical and modern classical schools in the 1980s and 1990s. In many ways, it was an amalgam of Keynesian and Friedman's ideas in macroeconomics, and led the way to the re-emergence of the classical doctrines. While its theoretical underpinnings have not survived, its impact

firms and unions over relative income shares. Still others claimed that firms followed a full-cost pricing policy, with the money supply accommodating itself to the resulting price level because the central bank did not want the unemployment rate to rise.

[31] Numerous applications of the St Louis equation showed that its empirical findings differed among countries, periods and the definitions of the policy variables. However, their basic conclusion, that money supply changes have a strong short-term impact on the economy, remained fairly robust.

on monetary policy has proved to be longer lasting. On this, its contributions were that money matters, that the control of the money supply is important for containing inflation and that the responsibility for inflation rests with the central bank.

Empirical evidence

One way of assessing the empirical validity of the implications of the classical models is by comparing their implications with the stylized facts set out early in this chapter. For the long run, the classical models imply that output and unemployment in the economy are independent of the money supply, price level and inflation, while the relationship between money supply and price level is proportionate. Both implications seem to be confirmed by data over long periods of time (Kormendi and Meguire, 1984; Geweke, 1986; McCandless and Weber, 1995; Taylor, 1996; Lucas, 1996), though some studies show a positive correlation between output and inflation in a general context of low inflation (McCandless and Weber, 1995) while others show a negative one, especially at higher inflation rates (Barro, 1996).

On the short-run or dynamic impact of an expansionary monetary policy, i.e. an expansionary money supply and/or decrease in interest rates, the modern classical model implies, through the Friedman–Lucas supply analysis, that prices and inflation will increase. If the increase is unanticipated, output will rise and unemployment will fall; if it is anticipated, there will be no change in output and unemployment. However, as the stylized facts show, the dynamic response of output and unemployment to an expansionary monetary policy is hump-shaped: output first increases (and unemployment falls) for several quarters, and then begins to decrease (Sims, 1992; Ball, 1993). This evidence contradicts the response pattern implied by the modern classical school. Further, while the impacts of anticipated and unanticipated monetary policy can sometimes be different, this difference is not always significant. In any case, an anticipated monetary policy usually does have a significant impact on output and unemployment, and unexpected price and inflation changes explain only a small fraction of the output changes that occur. Further, the real effects of monetary policy do not proceed through prior changes in prices and inflation rates, as asserted in the Friedman and Lucas supply rules (Lucas, 1996).

Therefore, the inescapable conclusion is that while the modern classical model does provide an acceptable long-run relationship between money and output/unemployment (i.e. money and monetary policy are neutral), it does not provide a satisfactory short-run theory of the impact of money and monetary policy on output and unemployment.

This chapter has presented the analyses of the classical paradigm in short-run macroeconomics. How does it perform empirically? For this, we rely on the assessment provided by Robert E. Lucas, who is associated with the modern classical school and has been a major contributor to it.

Lucas on the neutrality versus non-neutrality of money

In the “Nobel lecture” (1996), given on his receipt of the Nobel Prize in economics, Robert Lucas noted that:

> In summary, the prediction that prices respond proportionately to changes *in the long-run*, deduced by Hume in 1752 (and by many other theorists, by many different routes, since), has received ample – I would say decisive – confirmation, in data from many times and places….

> The observation that money changes induce output changes in the same direction receives confirmation in some data sets but is hard to see in others. Large-scale reductions in money growth can be associated with large-scale depressions or, if carried out in the form of a credible reform, with no depression at all.
>
> (Lucas, 1996, p. 668; italics added).

> *Sometimes*, as in the U.S. Great Depression, *reductions in money growth seem to have large effects on production and employment*. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so. Observations like these seem to imply that a theoretical framework such as the Keynes–Hicks–Modigliani IS/LM model, in which a single multiplier is applied to all money movements regardless of their source or predictability, is inadequate for practical purposes.
>
> (Lucas, 1994, p. 153; italics added).

Note that this quote states that changes in the money supply need not always be neutral and can (not must) have large real effects, lasting over a considerable period. But the modern versions of the classical paradigm do not provide a theory to explain such significant instances of the non-neutrality of money.

Lucas on the validity of the modern classical analysis for the short run

Lucas claims that "Macroeconomic models with realistic kinds of monetary non-neutralities do not yet exist." (Lucas, 1994, pp. 153–4). "… anticipated and unanticipated changes in money growth have very different effects" (1996, p. 679). However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that:

> *Only small fractions of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, they do not seem to be transmitted through price increases, as in Lucas (1972).*
>
> (Lucas, 1996, p. 679; italics added).

Note that this quote indicates that money supply changes do not have to first cause changes in prices before they affect real output. That is, output may change in response to monetary policy changes, even though the markets might not first or ever adjust prices. This implies that an expansionary monetary policy would produce output increases before being reflected in inflation (Mankiw, 2001). This is a powerful indictment of the long-run classical paradigm with its underlying assumption at both the microeconomic and macroeconomics levels, which is that following an increase in demand, prices are first adjusted by markets, and this is followed by adjustments in quantities demanded and supplied by economic agents.

Lucas on the state of macroeconomic theory

> Little can be said to be firmly established about the importance and nature of the real effects of monetary instability, at least for the U.S. in the postwar period. Though it is

widely agreed that we need economic theories that capture the non-neutral effects of

> money in an accurate and operational way, none of the many available candidates is without serious difficulties.
>
> (Lucas, 1994, p. 153; italics added).

Hence, given the above assessment by Lucas, it is fair to conclude that the short-run modern classical models, based on his model of price misperceptions in the commodity market and/or on Friedman’s errors in expectations occurring in wage contracts, fail to provide a satisfactory explanation of the empirical evidence on the impact of monetary policy on output, employment and unemployment. We must therefore continue the search for additional theories of such impact. The next chapter looks at various theories in the Keynesian paradigm, and their success in this objective.

15 The Keynesian paradigm

The Keynesian tradition differs from the classical one in not assuming that the economy is always in equilibrium with full employment or, if out of equilibrium, tends on its own to full employment within a reasonably short time. This was the core assertion of Keynes's *The General Theory*, published in 1936, and remains at the core of all models within the Keynesian tradition. This assertion has the corollary that appropriate macroeconomic policies could improve on the functioning of the economy. Most central banks do follow this recommendation.

The Keynesian model has been evolving ever since its basis was laid in Keynes's contributions in 1936. It has gone through many versions, with different versions taking centre stage at different stages in its evolution and many coexisting simultaneously. Its latest version is the new Keynesian model.

Key concepts introduced in this chapter

- ♦ Involuntary unemployment
- ♦ Phillips curve
- ♦ Demand-deficient Keynesian model
- ♦ NeoKeynesian model
- ♦ New Keynesian model
- ♦ Notional demand and supply functions
- ♦ Effective demand and supply functions
- ♦ Sticky prices
- ♦ Staggered wage contracts
- ♦ Implicit employment contracts
- ♦ Efficiency wages
- ♦ Taylor rule
- ♦ New Keynesian Phillips curve

This chapter reviews the Keynesian ideas on short-run macroeconomics. To start, the material on the Keynesian paradigm in Chapter 1 should be reviewed. That presentation argued that the Keynesian paradigm was the study of the pathology of the macroeconomy: it focuses on the causes, implications and policy prescriptions for the deviations of the economy from its Walrasian general equilibrium position and holds that the pursuit of appropriate monetary and fiscal policies can shorten the extent and/or duration of the deviation. Since there can be many causes of such deviations, their appropriate study requires not one unified model but many, some of which will be variations on the same theme, but there could also be models that are incompatible with one another. As such, there is no one model that can lay claim to be *the* Keynesian model. This chapter provides a small sample of the diversity of Keynesian models.

Since the classical model implies only transitory and self-correcting deviations from full employment, its emphasis tends to be on finding the long-run relationships. Since the Keynesian paradigm holds that deviations from full employment may not always be transitory and self-correcting, its focus is on finding the short-run dynamic relationships and the policies appropriate to them.[1] Keynes himself had expressed his strong disapproval of the long-run focus of classical macroeconomics in his time (and continuing at the present time) in:

> But this long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us that when the storm is long past the ocean is flat again.
>
> (Keynes, 1923, p. 80).

Keynesianism is a living tradition, evolving and refining its ideas over time, so that there are several versions of the Keynesian approach. The original version was Keynes's own ideas as set out in *The General Theory* (1936), followed by a number of evolving and quite diverse versions of the Keynesian framework,[2] representing a broad and often somewhat disparate set of ideas. While some of the earlier versions of Keynesianism are still part of its current mainstream, its most recent version is the new Keynesian (NK) one, which emerged only in the 1990s. Given this diversity, one needs to focus on the core themes in the Keynesian paradigm as a whole.

The common strands in the Keynesian approaches, broadly defined, are the reliance on some form of market imperfections and the angst over the performance of the economy, especially of the labor and commodity markets. Additional themes in the Keynesian literature, as in Keynes's *The General Theory*, are: the impact of changes in aggregate demand on output and employment, animal spirits (economic agents' psychology) and degree of confidence, market psychology, a consumption function that bases actual consumption on actual income through the multiplier, liquidity preference, presence of uncertainty and the possibility of default in financial markets (so that not all bonds can be assumed to be one-period riskless ones), moral hazard and contagion in financial markets, and macroeconomic instability rather than stability. While we touch on some of these topics at various places in this book, most are left to more specialized studies.

Keynesian models reject the perfectly competitive functioning of the labor market. Over time, given the complexity of this market, different Keynesian models have resorted

[1] Therefore, in empirical analysis using cointegration and error-correction techniques, the concern of the classical analysis is mainly with the properties of the cointegrating vector, while that of the Keynesian paradigm is mainly with the coefficients of the error-correction equation.

[2] There is considerable dispute as to whether any of the Keynesian models represents Keynes's own work. A close reading of Chapters 2 and 3 of *The General Theory* shows that they do not. It is therefore appropriate to make a distinction between the Keynesian models and Keynes's own analysis, though the former arose out of interpretations of the latter.

to different simplifying assumptions about it. At the risk of oversimplification, these were that:

(i) The nominal wage is fixed (1940s and early 1950s).
(ii) The nominal wage is variable but the supply of labor depends on the nominal and not the real wage (1950s and 1960s).
(iii) The structure of the labor market can be replaced by the Phillips curve (late 1950s and early 1960s) or the expectations-augmented Phillips curve (late 1960s).
(vi) The demand and supply of labor depend on the expected real (not nominal) wage, but the expectations on prices, needed to derive the expected real wage from the negotiated nominal wage, are subject to errors and asymmetric information between firms and workers (1970s and 1980s). Wage contracts are in nominal terms and staggered over time.
(v) The focus of the Keynesian analysis is on states other than full employment, so that the notional demand and supply functions of Walrasian and neoclassical economics are not applicable. The applicable concepts are those of effective demand and supply and the equilibrium – or "temporary equilibrium" as it is called by some writers – between them can occur at less than full employment and have different dynamic properties from those of neoclassical economics or its disequilibrium analysis. Such analysis is sometimes presented in the form of quantity-constrained models (1970s and1980s).
(vi) In the labor market, the real wage is an efficiency real wage that can be rigid in the short run. In addition, commodity prices are sticky. Labor markets also have implicit contracts and firm-specific skills. In the commodity market, prices are sticky because of menu costs (1980s and 1990s).
(vii) The economy has forward-looking, monopolistically competitive firms that optimize intertemporally, yielding staggered, discrete adjustment of individual commodity prices. In the aggregate, this results in an NK Phillips curve relationship with the price level adjusting more slowly than in a Walrasian economy with money neutrality. Further, on monetary policy, the forward-looking optimizing central bank follows a Taylor rule for setting its interest rate (after about the mid-1990s).

The intention of this chapter is not to go through each of the various versions of the Keynesian approach but to present just three versions to show the general pattern and variety of their reasoning and implications for monetary and fiscal policies. The common theme among these versions of the Keynesian ideas is that money is not neutral in the short run and that aggregate demand management can help to improve on its performance. The models that will be used to show this are the disequilibrium or temporary equilibrium (also called the demand-deficient) model, the Phillips curve model and the NK model. Our presentation will cover the following models and ideas:

(I) The demand and supply functions are as in the neoclassical model but the labor market does not always clear, or clear fast enough, in notional terms. Further, economic agents react faster than markets to disequilibrium once it has emerged. The models used for this presentation are the effective (deficient) demand ones with involuntary unemployment.
(II) The individual equations of the labor market are replaced by the Phillips curve as a mode of encapsulating the behavior of the labor market and variations in it.

(III) There are various aspects, such as staggered nominal wage contracts, implicit contracts, efficiency wages, menu costs, etc., that affect the functioning of the labor and commodity markets. These somewhat disparate ideas, as compared to an integrated macroeconomic model, can be grouped under the label of NeoKeynesian economics.

(IV) The final model presented is that of the integrated NK approach, encompassing an NK price adjustment process (the NK Phillips curve) and a Taylor rule to set the economy's real interest rate. This approach differs from the preceding ones by using a stochastic intertemporal general equilibrium model with monopolistically competitive firms to derive the price adjustment equations of the macroeconomic model.

Keynesian models omitted from our presentation

In particular, the models that are within the Keynesian tradition but have been *omitted* from this chapter are:

(a) Those that assume that the supply of labor depends in some way on the nominal wage and not merely on the real one. The extreme form of this hypothesis, that the nominal wage is rigid downwards, is also not presented. The latter was popular in the 1940s and 1950s and the former was common in the 1960s. Neither is now in vogue among Keynesian economists.

(b) The labor market clears but the supply of labor depends on the expected real wage, and price expectations are subject to imperfect and asymmetric information. Since such a model involves uncertainty and expectations, its presentation is to be found in the preceding chapter's analysis of the expectations-augmented Phillips curve (EAPC).

(c) Many Keynesians in the 1950s and the 1960s argued that inflation was due to the market power of monopolistic firms and/or labor unions. In some of these (cost-push) models, unions pushed for higher wages, firms followed a full-cost pricing policy and the central bank accommodated the money supply to the price level so as not to raise the unemployment rate.

On the product, money and bond markets, most Keynesians up to the mid-1990s seemed willing to accept the assumption that the central bank exogenously provides the money supply to the economy, as well as the IS–LM model's Keynesian–neoclassical synthesis of aggregate demand, which had evolved by the 1960s. The more recent (post-mid-1990s) new Keynesian models discard the LM curve, assuming that the central bank exogenously sets the interest rate or follows a Taylor rule on it, and combine it with an IS equation. The relevant IS–LM and IS–IRT models of aggregate demand in the open economy were presented in Chapter 13. This chapter takes their analysis of aggregate demand as given.

A caution on the categorization of Keynesian models

The following cautionary notes at this stage might provide some general guidance in comparing the neoclassical and Keynesian paradigms.

1 Keynesians in general accept the specification and conclusions of the classical paradigm on the long-run equilibrium of the economy. That is, in long-run equilibrium, the economy will have full employment and money neutrality. However, the Keynesian models differ from the classical paradigm in their conclusions about

the short-run and short-term functioning of the economy. In particular, they focus on those deviations from long-run equilibrium that are not necessarily transient and self-correcting.

2 It is often contended that the distinguishing difference between the classical and Keynesian paradigms is that the former assumes the flexibility of nominal wages and/or prices, while the latter assumes them to be rigid. As will be shown in this chapter, the rigidity of nominal wages and/or prices is not a necessary component of some of the Keynesian versions. This is not to say that the Keynesian models cannot be based on such an assumption; it is rather to assert that they do not necessarily require such an assumption and that the qualitatively distinct Keynesian policy results can be derived with flexible prices and nominal wages provided that the condition of instantaneous market clearance is not imposed.

3 Further, the Keynesian models do not necessarily need to rely upon irrational (such as price illusion) or myopic (one-period optimization) economic behavior but can derive their distinctive conclusions under the rational demand and supply behavior of economic agents, given the conditions that come about in the relevant markets.

4 It is sometimes contended that while the Keynesian models study disequilibrium in the economy, classical models study its equilibrium properties. However, while some Keynesian models do focus on disequilibrium, others, such as the NK model, impose the requirement of general equilibrium, in which all markets are assumed to clear.

5 It is also often contended that, while the (current) models of the classical paradigm are based on micro foundations with optimal behavior by economic agents, those of the Keynesian models are not so based. This is not always so. The current crop of new Keynesian models derives its macro relationships from micro foundations with optimizing behavior.

Sections 15.1 to 15.4 present three of the main versions of the Keynesian model. Section 15.5 derives the reduced-form Keynesian relationship between output and the money supply. Section 15.6 presents the empirical assessment of the validity of the Keynesian model.

Keynesian model I: models without efficient labor markets

A market is efficient if it instantly restores equilibrium following any shift in demand or supply. Conversely, an inefficient market is one that does not have instantaneous market clearance. In the context of the labor market, note that this market is really very many diverse markets, separated by skills, location, different firms, and so on, and often with implicit long-term contracts and insider–outsider trading, etc. These factors provide plenty of scope for those who want to assume that the labor market is not efficient, while leaving others to assume that it can be taken to be efficient, or approximately so, for macroeconomic analysis.

Keynes (1936) argued that in a monetary economy the worker is generally paid a wage rate that is not guaranteed in terms of its purchasing power but is, rather, a nominal wage rate. It is the nominal wage rate that is negotiated between an employee or his union and the employer. Once a nominal wage rate has been set in an explicit or implicit contract, under normal economic circumstances neither the employee nor his union seems willing to accept a cutback in the set wage rate. However, while workers are not willing to accept a cutback in nominal wage rates, they seem more willing to tolerate a reduction in the purchasing power of their nominal wage if this is brought about by changes in the purchasing power of money.[3]

[3] This is, however, a short-term phenomenon in the modern industrialized economy. Workers and their unions are nowadays sufficiently sophisticated to realize the losses in real income due to inflation and try to base the

A simple version of this ("rigid wage") model, in which the nominal wage rates are assumed to be rigid downwards, was an early version (1940s and 1950s) of the Keynesian model. This was supplanted in the 1950s and 1960s by a "nominal wage model" in which the supply of labor depended on the nominal wage rate, which was flexible and determined by the labor market.[4] However, these two models are not now seriously included among the models of the Keynesian paradigm and are omitted from this edition. Those still interested in their exposition can find the latter in Handa (2000, Ch. 15) or Patinkin (1965, Ch. 13). However, we note in passing that the assumptions of this nominal-wage model did not include the rigidity of nominal wages and prices, or of real wages, nor did it imply it. Further, this model assumed market clearance, but at the equilibrium nominal, not real, wage.

The assumption of labor supply depending on the nominal rather than the real wage implies money illusion or myopia by labor, which is empirically not valid over any extended period of time. Many Keynesians and especially post-Keynesians have argued that Keynes's *The General Theory* did not make this assumption but agreed with the neoclassical assumption that workers would base the supply of labor on the purchasing power of the nominal wage, i.e. on the real wage rate. They also argue that Keynes assumed that the modern economy, with numerous industries and firms and with decentralized wage negotiations, does not possess a mechanism that would ensure that the labor markets are normally in equilibrium at full employment with an equilibrium real wage. Hence, the claim being made under this argument is that the major distinction between the Keynesian and the neoclassical models is that, while the neoclassical model makes the assumption of labor market equilibrium as the usual state, Keynes did not do so. That is, Keynesian models of this type specify the labor market demand and supply functions as:

$$n^{\mathrm{d}} = n^{\mathrm{d}}(w) \qquad \partial n^{\mathrm{d}}/\partial w < 0 \tag{1}$$

$$n^{\mathrm{s}} = n^{\mathrm{s}}(w) \qquad \partial n^{\mathrm{s}}/\partial w > 0 \tag{2}$$

These behavioral functions are identical with those in the neoclassical model. However, on the basis of the empirical belief that the labor markets are usually not in long-run equilibrium, there is no assumption of labor market equilibrium in this type of Keynesian model. Since firms are assumed to be able to hire the amount of labor that they demand, the market clearance condition ($n = n^{\mathrm{f}} = n^{\mathrm{d}} = n^{\mathrm{s}}$) of the neoclassical model is replaced by:

$$n = n^{\mathrm{d}} \leq n^{\mathrm{s}} \tag{3}$$

Note that (3) does not assume that labor market equilibrium at full employment will never exist in the economy. This equation implicitly assumes that the modern capitalist economy does not possess sufficient mechanisms to ensure continuous full-employment equilibrium or achieve it within a reasonable period of time after a shock. Further, such an

negotiated wage on the expected rate of inflation. If the actual inflation rate is higher than the expected one, workers try to get compensation for it at the next round of wage negotiations, so that in the long run labor supply will be effectively based on the real wage rather than on nominal wages. We will return to these considerations at a later stage in this chapter.

[4] The above is clearly a rather simplified picture of the wage bargain. A somewhat better picture emerges if it is assumed that workers supply labor on the basis of the expected real wage rate and the expectations are explicitly modeled. This is particularly so when nominal wage rates and prices are rising and it is difficult for

workers to perceive the actual change in the real wage rate.

economy can get stuck at a level of employment below full employment, and these levels can also be equilibrium states.[5] That is, the Keynesian models allow the possibility of multiple macroeconomic equilibria, each with a different level of output and unemployment. One of these equilibria is that of full employment, so that we have to distinguish between the full-employment equilibrium and other equilibria with less than full employment.

The justification for involuntary unemployment

Designate the *long-run equilibrium* (*full-employment)* level of employment as n^f, and the rate of unemployment consistent with it as u^n u^n is the natural or Walrasian (full-employment) equilibrium rate of unemployment consistent with the structure of the economy, including any wage rigidities such as specific skills, minimum wage laws, labor unions and work preferences. Also designate the *short-run equilibrium* unemployment rate as u^*, which differs from u^n due to errors in price expectations and the costs of adjusting prices, employment, output, etc. Separate the two components of u^* due to expectations errors, which is an element of the short-run modern classical model (see Chapter 14), and those due to adjustment costs, some of which underlie many Keynesian models (see Section 15.3), as u^{J*} and u^{JJ*}, respectively with u^* u^{J*} u^{JJ*}.

Note the identity. $=$ $+$

$$u \equiv u^n + (u* - u^n) + (u - u*) \tag{4}$$

If there are no errors in price expectations and no adjustment costs, u^* (short-run equilibrium unemployment rate) will be identical with u^n (long-run equilibrium unemployment rate). Since the actual economy cannot be expected to be always and at every instant in equilibrium, or, as Keynes would have claimed, in equilibrium most of the time, (u u^*) would not always equal zero. Define u^i u u^*, where u^i is the deviation of unemployment from its short-run equilibrium value. With u^n (u^* u^n) as the short-run equilibrium unemployment rate, u^i will be indicative of the failure of the economy to be in either short-run or long-run equilibrium. u^i is usually labeled the *involuntary unemployment rate*. Its value can be positive, zero or negative, with a negative value likely to occur near the peak and a positive value near the trough of the business cycle. From the perspective of the actual economy and policy, u^i would be zero only if the economy, not the model, is operating in equilibrium. From the perspective of a model, u^i would be zero in the solution of the model if the analytical condition of short-run equilibrium is imposed in deriving the solution. Keynes and the Keynesians (other than the new Keynesians) used the former perspective and argued that the actual economy is usually not in equilibrium. Further, given their focus on deficient demand rather than on excess demand, their analysis implied positive involuntary unemployment when demand is deficient relative to the short-run equilibrium output level.

Keynesians argue, therefore, that the authorities should keep a close watch on the economy and, when there is significant involuntary unemployment due to a deficiency in aggregate demand, they should use monetary and/or fiscal policies to increase demand by an appropriate amount. If they succeed, the economy will eliminate such involuntary unemployment and perform at its full employment potential.

[5] The term "equilibrium" is defined here as a state from which there is no inherent tendency to change. It specifies the values of the endogenous variables implied by the model for a specified set of values of the exogenous variables in the model.

The class of models that fit the preceding remarks are often called "*deficient-demand models*," though they allow the absence of deficient demand in the limiting case of full employment. The following subsection presents an example of a deficient-demand analysis, noting that there can also be other versions that fall into this category.

Keynesian deficient-demand model: quantity-constrained analysis

Keynes had argued that the economy may not always generate aggregate demand equal to the full-employment supply of commodities. The former comes from the decisions of very many households, firms and the government (and exports in the open economy) in the form of consumption, investment and government expenditures. The latter is generated by firms, which undertake production to meet expected sales. There is no a priori reason for the two to be equal. The two could, however, be brought into equality by efficient markets that adjust the price to the equilibrium level. However, Keynes argued that the markets were not efficient (i.e. following a shift in demand or supply, adjusting the price instantly to equate demand and supply), so that firms and households would make their production and consumption decisions in response to expected demand and incomes, rather than to changes in the market price. In brief, markets were sluggish in their adjustment, and firms and households reacted faster than markets. This implies that firms would base their output and investment on expected demand, so that changes in aggregate demand would change aggregate output and employment. Households, in deciding on their consumption expenditures, reacted to their expected income and job prospects and not merely to their real wage rate, so that their consumption, and therefore aggregate demand, would respond to their job prospects. Hence, a fall in aggregate demand would reduce output and employment, which would, in turn, worsen incomes and job prospects and lower consumption (Clower, 1965; Patinkin, 1965; Leijonhufvud, 1967, 1968).

Writing in the Great Depression, Keynes argued that the economy often functions below its full-employment level. While a depression has not been experienced since the 1930s by the major economies, all economies do suffer periodic recessions. Keynes's factual statement is generally taken to be correct for recessions.[6] The most plausible reason for output falling below its full-employment level is a demand deficiency originating with a fall in aggregate demand. To show this, start with the initial state of full-employment equilibrium in the economy and assume that aggregate demand falls for some reason, thereby creating deficient demand relative to the full-employment level. Following the ideas of the preceding section, assume that the labor market does not instantly clear, so that some of the workers become involuntarily unemployed because of the fall in aggregate demand. These workers will not receive the wage they would have got if they had been able to sell their labor according to their supply curve. Their lack of income forces them to reduce their demand for commodities and real balances below that specified by their Walrasian functions as derived in Chapter 3 and the neoclassical functions specified in Chapter 14. Hence, these latter functions are not "effective" – that is, operational – in the state of involuntary unemployment. They can only be effective if there is full employment in the economy and workers could sell all the labor they wanted to at the existing wage rate.

[6] It is a well-established stylized fact that virtually all recessions have a fall in aggregate demand and a fall in output, as well as a fall in the inflation rate.

Demand and supply functions derived under the assumption of the simultaneous clearance of all markets are called *notional functions*.[7] They are the ones derived in Walrasian analysis and used in neoclassical models. Since involuntary unemployment means that at least one of the markets does not clear, the use of notional functions to analyze the existence or non-existence of involuntary unemployment begs the question and is inappropriate. The more appropriate analysis would be to posit that the real-world – that is, actual – demand and supply functions are approximated by effective functions, of which the limiting case is notional functions.

Effective functions for any market that take account of the non-clearance of other markets are also called *Clower* or *quantity-constrained functions* and the macroeconomic analysis based on them is similarly called quantity-constrained analysis. Such analysis clearly belongs among the Keynesian stable of macroeconomic models and became popular during the 1970s and 1980s.

Quantity-constrained analysis can encompass the possibility that any or all of the four markets of the macroeconomic models need not clear instantaneously. However, such Keynesian analysis focuses on non-clearance in the commodities[8] and labor markets. In particular, the main initial impulse of such models is a fall in the aggregate demand for commodities due to a fall in investment, in consumption, in government expenditures or exports, or in money supply.

Dynamic analysis following a fall in the aggregate demand for commodities

While the reader is referred to an appropriate macroeconomics book for a proper treatment of effective demand models, we pursue here Patinkin's (1965, Ch. 13)[9] application of this analysis to the market for labor in order to derive the role of monetary policy in such models. Assume that the demand and supply functions for labor are the neoclassical notional ones, as shown in Figure 15.1b, and that initially the economy is at full employment n^f in Figure 15.1b and full-employment output y^f in Figure 15.1a. Now assume that a shock reduces aggregate demand to $y^d{}_0$ so that a demand deficiency emerges in the economy such that the firms are not able to sell the full-employment output y^f at the existing price level.[10] The actual aggregate demand $y^d{}_0$ can be supplied by the employment of $n^d{}_0$ workers. In Figure 15.1b, the marginal product of labor for $n^d{}_0$ workers is MPL_0, which is above the full-employment wage of w^f. However, if firms were to employ more than $n^d{}_0$ workers, they would not be able to sell

[7] See Chapter 18 on Walras's Law for additional exposition on these functions.

[8] Such inequality of demand and supply occurs in notional terms, whereas for the commodity market the equality of the actual or effective demand and supply would still occur and determine prices. For the labor market, such non-clearance means that the notional supply of labor exceeds the notional demand.

[9] Patinkin was one of the most distinguished contributors to the neoclassical approach, even though his contribution on the following analysis of deficient demand was in the spirit of Keynesian economics.

[10] Neoclassical reasoning at this point would be that the aggregate price level would fall, creating a real balance effect which shifts the LM curve to the right, thereby increasing incomes and inducing an increase in consumption spending, so that the fall in aggregate demand would be reversed. Keynesians claim that this process does not work or is too slow, for many reasons: uncertainty of whether the fall in demand is transitory or longer lasting; firms may find it optimal not to change their price lists immediately; the real balance effect is quite ineffective and extremely slow in increasing demand; etc. Hence, Keynesians tend to assume that, for analytical realism, it is better to assume constant prices rather than falling prices, so that the response of firms to the fall in demand is to adjust the quantity produced rather than to reduce prices. The following analysis follows this Keynesian procedure.

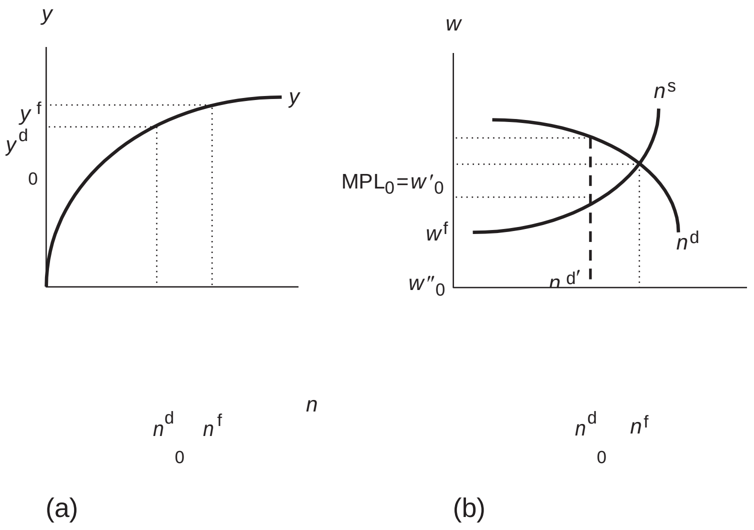


Figure 15.1

the extra output so that their *marginal revenue product* would be zero. Hence, if aggregate demand fell to y^{d}_{0}, firms would cut employment to only n^{d}_{0} workers.[11]

Extending this argument to Figure 15.1b, the employed n^{d}_{0} workers can be paid nominal wage rates which can change, as can the price level, with the resultant real wage being anywhere within the range w^{J}_{0} and w^{JJ}_{0}, without a change in the firms' employment of n^{d}_{0} workers, so that real wage rates could drift up or down from the initial equilibrium level of w^{f}.[12] Hence, the decrease in employment (from n^{f} to n^{d}_{0}) can be accompanied by either an increase or a decrease in the real wage rate of the employed workers. Wages may therefore follow a pro-cyclical or counter-cyclical pattern: some recessions and some parts of a given recession could show wages falling while others show them to be rising. If wages rise, it could be claimed that the rise in wages is the cause of falling employment, when this rise is itself only an effect while the true cause was the initial fall in aggregate demand.

The above effects are only partial or initial ones. Since the unemployed workers do not receive any income, they cut back on their consumption demand. Further, if wages were cut below w^{f}, the lower incomes of the employed would also lead to a reduction in consumption.[13] The consequent fall in aggregate demand further reduces the effective demand for labor[14] and acerbates the recessionary effects derived in the preceding paragraph.

The essential components of the preceding analysis are:

(a) The economy's industrial and market structures do not lead to instant market clearance in all markets, so that rational firms cannot assume this to be so, and must react in their

[11] As against this dynamic reaction, neoclassical economics claims that firms will cut prices, not output, in response to a fall in aggregate demand; or that markets will act fast enough to deliver the lower prices for all commodities sufficiently to allow firms to sell all they want before individual firms react by cutting their

production. Intuitive knowledge of the economy tends to favor the Keynesian response pattern.

12 Real wages will rise if the price level falls faster than nominal wages, they will fall if the price level falls more slowly than nominal wages, and will stay constant if both prices and wages fall in the same proportion.

13 The fall in employment usually increases, among the employed workers, the subjective risk of staying employed, so that such workers also tend to cut back on consumption in order to increase precautionary saving to provide for the eventuality of becoming unemployed.

14 The new demand curve is not the neoclassical notional one n^d but an effective one, and one cannot proceed with analysis using n^d.

production and labor demand strategy to emerging demand conditions for their products as they see fit. They react faster than markets to changes in demand for products and labor.

(b) There is no coordinating mechanism or coordinating agency (such as the Walrasian auctioneer with costless and open recontracting while adjustments to a new equilibrium are going on) in the economy which will work out the equilibrium wages, employment, output and prices instantly following shifts in demand and supply. Markets are competitive and tâtonnement towards equilibrium values may occur, but takes time, during which firms and workers make decisions on employment.

(c) There is no nominal wage or price rigidity. Prices and wages would be market determined if the markets were efficient or at least adjusted faster than economic agents in response to shocks.

(d) Firms and households are rational and react to the conditions that they face in the absence of perfect markets.

The lack of perfect market structures ensuring instantaneous adjustment to equilibrium or the lack of a coordinating mechanism or agency – with firms and workers not waiting out this process but responding faster than the sluggish markets to a fall in demand is the critical reason in effective demand models why an economy in which a demand deficiency has emerged need not rapidly move to the full-employment level. Critical to the dynamic responses of economic agents are their rationally expected values (of aggregate demand, prices and employment) that apply in the period (day, month, quarter, etc.) ahead. Their estimates of these expectations are reflected in some way by economic/business analysts' indices of business and consumer confidence.

It seems reasonable to posit that a very mild fall in aggregate demand in an overall time path of full employment can leave both business and consumer confidence intact and not produce a reduction in output and employment, whereas a more significant one or/and over a longer period, especially when backed by past experienced recessions, would produce a loss in such confidence and take the economy onto a dynamic path to involuntary unemployment levels. The irony of this remark is that the deliberate pursuit of aggressive Keynesian policies to maintain aggregate demand at the full-employment level leads to a dynamic response by consumers and firms which maintains full employment; but a policy of leaving the economy alone, as the classical economists recommend, brings into play dynamic responses which take the economy away from full employment. In support of this contention, the Western economies experienced very shallow and short recessions during the Keynesian period of the 1950s and 1960s, whereas the recessions lengthened and worsened with the resurgence of classical economics in the 1980s and 1990s.

Note that new Keynesian models of the post-1990 variety replace the assumption that economic agents react faster than sluggish markets by the assumption that firms are price setters because they are in monopolistic competition.

Optimal monetary policy in the Keynesian demand-deficient economy

To derive the optimal role and impact of monetary and fiscal policies in such a context, assume that the economy is now at $n^d{}_0$ of employment and $y^d{}_0$ of output in Figures 15.1a and 15.1b, and suppose the monetary authorities pursue an expansionary monetary policy. This increases aggregate demand in the economy, thereby shifting the n^d curve from n^{d^J} towards n^f and increasing output beyond $y^d{}_0$. Since output increases in response to the increase in aggregate demand, prices may or may not increase. The increase in prices will depend on how deficient the earlier demand was and how large was the earlier excess capacity in the economy, but, in any case, prices will not rise in proportion to the increase in the money supply.[15] The expansionary monetary policy would have succeeded in increasing output in the economy. But once the economy has reached y^f – that is, there no longer exists any demand deficiency further expansions in the money supply will not increase output but merely cause a proportionate increase in the price level. This limiting case of demand-deficient analysis is, of course, the neoclassical full-employment case, in accord with the Keynesians' claim that their analysis is the more general one and encompasses the neoclassical full-employment and classical cases as a limiting case.

These arguments imply that there is no straightforward or linear relationship between increases in the money supply and real output, or between the rate of inflation and the level of unemployment. These relationships depend upon the state of the economy and the extent of the monetary expansion. Further, the transmission of the impact of money supply increases on output does not always require or go through price level increases.

The economy's responses to excess demand

We have so far considered the dynamics of a decline in demand from a full-employment level. But suppose aggregate demand increases when there already exists full employment and the economy starts with the natural rate of employment. The dynamic response patterns of firms and workers should be basically consistent, though in opposite directions, in the two situations since these would be based on their rational response behavior patterns, but the economy's constraints would be quite different between these cases. On the former, the individual firm, seeing an increase in the demand for its product, would tend to increase production through increases in employment, overtime worked, increased effort of employees and increases in efficiency. Each of these is feasible in the short run, with the increase in employment, beyond an initial full-employment level, occurring through increased working hours of part-time workers, through students delaying resumption of studies and through increases in the overtime put in by employed workers, etc.[16] While such increases in employment and output above their full-employment levels can and do occur, as long drawn-out booms indicate, their scope is constrained by the economy's short-run flexibility for these variables. Hence, while the increases in aggregate demand can and do increase output and employment in the short run beyond their full employment levels – and the economy can go below its natural rate of unemployment – the increase in prices is more likely, and faster, than the fall in prices in response to a decline in demand from the full-employment level.

Most central banks now believe that changes in aggregate demand do produce changes in the economy's output and that the economy sometimes operates below full employment and sometimes above it. Evidence of these is provided by the current popularity of the Taylor rule, which tries to limit the gap between actual output and full-employment output, as well as the deviation of inflation from a target level, by the impact of monetary policy on aggregate demand.

[15] The real wage may rise or fall, depending upon where it was earlier in relation to w^f.

[16] However, they cannot be sustained over time (e.g. workers get tired of putting in undesired overtime) and such increases in employment cannot be assumed for the long run.

The eclipse of the deficient-demand analysis

The above deficient-demand analysis was bypassed by new developments, especially the new Keynesian modeling, in the Keynesian paradigm during the last quarter of the twentieth century. One reason for this was that its theoretical development had reached a dead end. Further, it was not intellectually challenging in the context of precise mathematical modeling, especially of macroeconomic dynamics, which had come to dominate macroeconomics. An additional reason was the adoption by the new Keynesians of the intertemporal general equilibrium methodology under imperfect competition, which, by the nature of its assumption of general equilibrium, excluded the dynamic effects of deficient demand. However, this intellectual shunting aside of deficient-demand analysis does not necessarily affect its validity, so that it remains a component analysis of the Keynesian paradigm.

Keynesian model II: Phillips curve analysis

Phillips curve

In 1958, A.W. Phillips, on the basis of statistical observations for the UK proposed a negative relationship between the rate of nominal wage growth and the rate of unemployment. This was subsequently extended to show a negative relationship between the rate of inflation and the rate of unemployment, with the name “Phillips curve” being attached to both of these relationships. During the late 1950s and the 1960s, Keynesian economics came to embrace this curve, incorporating it in preference to a structural specification of the labor market, as in equations (1) to (3).

Phillips (1958) had plotted the rate of change of nominal wages against the rate of unemployment for the UK over several periods from 1861 to 1957, and found that the data showed a downward-sloping curve. That is, the plotted relationship was of the form:

$$\overset{\circ}{W} = f(u) \tag{5}$$

where $\overset{\circ}{W}$ is the rate of increase of the nominal wage rate and $f^{J} < 0$. One explanation for (5) is that unemployment represents the degree of labor market tightness, so that the higher the level of unemployment, the smaller will be the increase in the nominal wage.

Equation (5) soon evolved into its inverse and then into a relationship of the form:

$$u = g(\pi) \tag{6}$$

where $g^{J} < 0$. This relationship is drawn as the PC curve in Figures 15.2a and 15.2b.

The transition from (5) to (6) comes from the link between the nominal wage rate and inflation: nominal wages represent the main element of the cost of production, so that an increase in nominal wages will induce firms to increase their prices; alternatively, an increase in prices causes labor to ask for compensation in the form of wage increases. Hence, there is a positive relationship between $\overset{\circ}{W}$ and π, which, when substituted into (5), yields (6).

The estimated forms of the Phillips curves proved to be convex to the origin. To explain this curvature, it was argued that the response of nominal wages to excess demand was non-linear, and that decreases in unemployment caused successively greater increases in nominal wages. Further, a decrease in employment induces a smaller decline in wages than the increase in wages brought out by a corresponding increase in job vacancies, so that increases

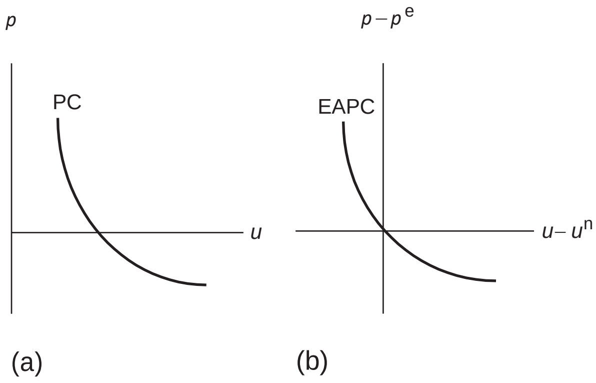


Figure 15.2

in employment in some industries with corresponding decreases in others will bring about a net increase in the average nominal wage rate. Hence, both the level of unemployment and its variance among industries together determined the Phillips curve relationship.

For the pre-1970s data, numerous studies for many countries, including Canada and the USA, seemed to confirm the validity of the Phillips curve. Even though the relationship seemed to differ between periods and countries, the general form of the relationship seemed to be valid for the 1950s and 1960s and became a mainstay of many Keynesian models, replacing the labor market structural relationships, during the 1960s and 1970s. As a consequence, many Keynesian economists in the 1960s and early 1970s assumed the Phillips curve to be stable and recommended its use as a trade off between inflation and unemployment by the monetary and fiscal authorities, calling on them to use their policies to change aggregate demand to achieve the inflation rate specified by the Phillips curve as a concomitant of the desired rate of unemployment in the economy. In this sense, while (6) was a constraint on policy choices, the Keynesians interpreted it as giving the authorities control over the unemployment rate in the economy, with the accompanying rate of inflation being an undesirable cost of the chosen unemployment rate. The Phillips curve provided the economic tool to support the economic philosophy of the Keynesians in the 1960s and 1970s that the monetary and fiscal authorities should try to achieve better levels of output than the economy would generate on its own, even though doing so would mean higher inflation rates.[17]

The expansionary monetary and fiscal policies resulting from central banks' attempt to take advantage of the Phillips curve tradeoff did lead to rising rates of inflation.[18] Once the inflation reached unacceptable levels, with a momentum towards further increases, the central banks would resort to monetary stringency to fight it. However, by this stage, the public had come to expect higher inflation and the inflation-fighting monetary contractions resulted mostly in increases in unemployment, reaction to which could force the central bank to again

[17] There was also at that time considerable skepticism among the Keynesians on whether the central bank could control inflation, since it was often attributed to cost-push forces, or because it could only control the money supply, which was only a small part of liquidity in the economy. The latter point was discussed in Chapter 2 under the topic of the Radcliffe report.

[18] Boschen and Weise (2003) explore the origins of 73 inflation episodes in OECD countries from the 1960s to the 1980s. They report that inflation often started because, out of a belief in the short-run Phillips curve,

central banks sought the short-term benefits of higher growth without considering the costs of disinflation later.

pursue a monetary expansion, thereby producing a “stop–go” policy pattern. The role of expectations proved vital to the impact of changes in inflation on output and unemployment, and led to the modification of the Phillips curve to the expectations-augmented curve.

The expectations-augmented Phillips curve[19]

The Phillips curve was challenged by Milton Friedman and the monetarists in the 1960s and 1970s. They argued that if inflation were perfectly anticipated, the labor contracts would reflect it so that the nominal wage would increase by the expected rate of inflation. Consequently, the expected inflation rate would not affect the real wage rate, employment or output. Hence, at the expected inflation rate, the unemployment rate would be the natural rate so that only the unanticipated rate of inflation would cause deviations in the actual from the natural rate by reducing the real cost of labor and other inputs. That is, according to Friedman, the proper relationship between u and π is not (6) but has the form:

$$(u^* - u^{\mathrm{n}}) = f(\pi - \pi^{\mathrm{e}}) \tag{7}$$

with $f^{J} < 0$ and $f(0)$ 0. u^{n} is the long-run equilibrium unemployment rate, while u^* is the short-run equilibrium unemployment rate in the presence of errors in inflationary expectations. In the limiting case of expectational equilibrium, an aspect of the “long run,” π π^{e} and $u^* = u^{\mathrm{n}}$. (7) is the expectations-augmented Phillips curve, as shown as the curve EAPC in Figure 15.2b. Its analysis and policy implications were presented in Chapter 14.

Empirical research and the widespread experience of stagflation in the mid and late 1970s in the industrialized economies seemed to show that (6) was unstable in a period of accelerating inflation. Further, (7) seemed to perform much better in such periods, especially at high and accelerating rates of inflation. In response to this analysis and evidence, Keynesian models after the 1970s tended to drop the Phillips curve as a primitive element of their models, though most such models can generate some form of it as an implication. Its most frequent resurgence is in the form of the new Keynesian Phillips curve, which is rather more in the nature of firms’ output–price adjustment equation than a true Phillips curve based on labor market behavior.

The relationship between the Phillips curve and the expectations-augmented Phillips curve

To look at the relationship between the actual Phillips curve and the expectations-augmented Phillips curve, start with the identity:

$$u \equiv u^{\mathrm{n}} + (u^* - u^{\mathrm{n}}) + (u - u^*)$$

where u is the actual unemployment rate, u^* is the short-run unemployment rate in the presence of errors in price expectations and u^{n} is the long-run equilibrium unemployment rate. From the Friedman and Lucas analyses in Chapter 14, we have:

$$u^* - u^{\mathrm{n}} = f(\pi - \pi^{\mathrm{e}})$$

[19] The formal derivation of this curve is presented in Chapter 14, which discusses uncertainty and expectations.

Therefore,

$$u = u^{\text{n}} + f(\pi - \pi^{\text{e}}) + (u - u*) \tag{8}$$

The Phillips curve focuses on the determinants of ($u\ \ u^{\text{n}}$). There could be many determinants of this difference. One of these relies on errors in price or inflationary expectations. Keynesian models provide several other determinants, with only one of them being the demand-deficient analysis. New Keynesian models, presented later in this chapter, provide another form of the Phillips curve, which arises from monopolistic competitive behavior in commodity markets and price adjustment costs.

To emphasize the rarely recognized point about (8), the expectations-augmented Phillips curve is a component of the Phillips curve analysis, but may not even be the most significant part empirically. However, in view of this relationship between the curves, the Phillips curve does shift if there is a change in expectations of inflation, but this is only a part of the story.

The Phillips curve as specified by (6) or (8) is also different from the new Keynesian Phillips curve discussed later in this chapter.

Components of neoKeynesian economics

We group under this heading the economic theories developed since the 1970s to provide a foundation for the Keynesian tenets that involuntary unemployment can exist in the economy and that changes in aggregate demand in the economy can affect output, at least in the short run. These theories have in many ways supplanted the traditional Keynesian (nominal wage, demand-deficient and the original Phillips curve) models presented in the earlier sections of this chapter. We will present three of these theories. They are the efficiency wage theory leading to real wage rigidity, a theory of rigid or sticky prices based on sluggish price adjustments, and a theory of implicit contracts leading to labor hoarding. These are used in some combination to derive the neoKeynesian conclusions that monetary policy, through aggregate demand, affects output and is not neutral in the short run, though it is neutral in the long run.

Efficiency wage theory

While the neoclassical and Keynesian theories presented so far in this book have taken the effort put in by each worker on the job to be constant, the efficiency wage model proposed by Akerlof (Yellen, 1984; Shapiro and Stiglitz, 1984) assumes that this effort is a function of the worker's wage. It can also be a function of other variables such as the unemployment rate and the unemployment benefits that also affect the opportunity cost of the job. In order to accommodate the concept of a variable effort on the part of workers, the firm's production function is modified from the usual one of:

$$y = f(n, \underline{K}) \qquad f_n > 0,\ f_{nn} < 0$$

to:

$$y = f(e(w)n, \underline{K}) \qquad f_e,\ f_n > 0,\ f_{ee},\ f_{nn} < 0,\ \ e_w > 0, e_{ww} < 0 \tag{9}$$

where *e* designates "effort," taking this to be a measurable variable. (9) keeps the capital stock constant for short-run analysis. Effort *e* is a function of the real wage rate *w*. Paying the workers more than the market clearing wage tends to increase their productivity by reducing shirking by workers, reducing turnover of workers, improving the average quality of job applicants and improving morale in the firm.

Focusing only on the shirking element, most jobs do not rigidly force the workers to work at a pre-set pace with a pre-specified productivity but allow them some leeway in their level of performance. Workers could, therefore, “shirk” on the job, thereby lowering their productivity or requiring the firm to prevent shirking through performance monitoring by “inspectors,” which imposes additional costs on the firm. If the worker is paid the market wage only, and is fired if caught shirking, he could obtain another job at the same wage and therefore would only lose by the extent of his search costs. But if he is paid a wage higher than the market wage, he has an incentive not to shirk and thereby lose a job with a better wage than he can get if he shirked and was fired for doing so. The absence of shirking, in turn, increases the worker’s effort and productivity. The firm therefore has an incentive to pay its workers more than the market clearing wage. For optimization, the wage paid will be that which yields the lowest labor cost per efficiency unit – that is, with labor measured in terms of its efficiency. Designate this optimal wage, known as the efficiency wage, as w^*. The profit-maximizing firm will then employ labor up to the point at which its marginal product equals its real wage, that is, by:

$$\partial y/\partial n = e(w^*) f'(e(w^*)n, \underline{K}) = w^* \qquad f'(e(w^*)n, \underline{K}) = \partial f/\partial(e(w^*)n) > 0 \qquad (10)$$

In equilibrium, all firms would pay the real wage w^*, assuming it to be greater than labor’s reservation wage.

The efficiency wage models assume the neoclassical labor demand and supply functions, with both demand and supply being functions of the real wage, so that the labor market is represented by Figure 15.3. At the wage w^*, higher than the market clearing wage, $n^d{}_0 < n^f < n^s{}_0$, so that employment at n^* is less than full employment and there exists involuntary unemployment equal to $(n^s – n^d)$. These unemployed workers are willing to accept the wage w^* or w^f or even somewhat lower wages but their competition for jobs will not reduce the market wage, since such a lower wage will be below the efficiency wage w^* and reduce the productivity of firms’ existing employees and their profits. Consequently, the labor market will maintain involuntary unemployment equal to $(n^s\ n^d)$ in the long run. Since such involuntary unemployment is a long-run feature of the labor market, it can be called the long-run involuntary unemployment. Because of its long-run nature,

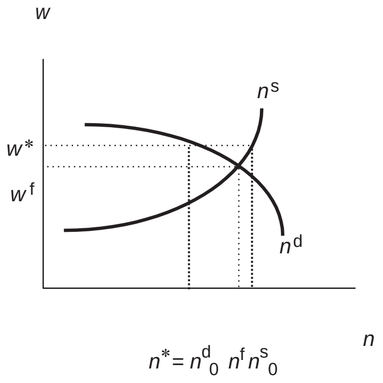


Figure 15.3

some economists propose its incorporation into the notion of structural unemployment and redefine the natural rate to encompass it. However, since the determinants of structural unemployment and of such long-run involuntary unemployment are quite different, we prefer to keep them as separate concepts. Nor do we merge the latter into the classical concept of the natural rate of unemployment. In the efficiency wage context, the long-run rate of unemployment will, therefore, be the classical natural rate plus efficiency-based long-run involuntary unemployment.

To explain the effects of a fall in demand on employment, rewrite (10) as:

$$p_i e(w^*) f^{J}(e(w^*)n, \underline{K}) = Pw^* = W^* \tag{11}$$

where p_i is the price of the firm's product, P is the price level and W^* is the nominal wage equal to Pw^*, with w^* still the efficiency real wage. (11) can be rewritten as:

$$e(w^*) f^{J}(e(w^*)n, \underline{K}) = (P/p_i)w^* \tag{12}$$

Now assume that there is a decline in the demand for the firm's product such that its relative price (p_i/P) falls. This will not change the firm's efficiency real wage w^*, but its demand curve for labor will shift down and its employment will fall. Conversely, an increase in the relative price of the firm's product will leave the efficiency real wage unchanged but increase its demand for labor and employment.

Now suppose that all product prices increase proportionately, so that we can use either (10) or (12). These equations imply that a change in the price level will not bring about a change in the efficiency wage w^* or the equilibrium values of employment n^* or involuntary unemployment u^{i*} in the economy. That is, this version of the efficiency wage theory does not imply that a change in aggregate demand – due to monetary or fiscal policy or other exogenous changes – will change aggregate employment and output. For this, we need to bring in the neoKeynesian theory of price stickiness, presented in the next subsection.

The efficiency wage theory implies long-run involuntary unemployment in the economy.[20] It can also be used to buttress the Keynesian claim that, following a fall in demand, the economy could move to an equilibrium with a still higher level of unemployment. For this, it is preferable to modify the effort function to a more realistic one as $e = e(w, u)$, where $\partial e/\partial w$ 0$\geq$and $\partial e/\partial u$ 0, so$\geq$that as unemployment rises, the firm can lower the efficiency wage necessary to get the maximum effort from the workers. In this case, a tradeoff will exist between the optimal efficiency wage and unemployment in the economy. A given fall in aggregate demand will, then, imply an equilibrium with a higher level of unemployment and a lower efficiency wage.

Costs of adjusting employment: implicit contracts and labor hoarding

The neoKeynesians also argue that it is optimal for firms and workers to enter into long-term implicit and explicit employment contracts. Such contracts are optimal for the firm because of the cost of hiring and training workers and the firm-specificity of skills acquired through training and learning on the job, so that the productivity of such a skilled worker will be

[20] It also explains the existence of dual markets, wage distributions among workers with identical skills and certain types of wage and job discrimination.

greater than of new hires. The worker also benefits from this higher productivity through higher wages in his existing firm than if he were to quit and join other firms. This mutual benefit from continued employment implies that the firm will try to retain its skilled workers if it can do so through a period of reduced demand for its output, rather than laying them off immediately. The firm therefore finds it optimal to lay off fewer workers than is justified by the fall in demand, leading to a form of labor hoarding during recessions (Okun, 1981). Such hoarded labor works less hard during recessions because there is less work to do, or is often diverted to such tasks as maintenance. If a worker is laid off, he also has an incentive to wait to be recalled by his old employer rather than immediately accept a job with another firm in which his productivity and wage will be lower. Hence, reductions in aggregate demand in the short run partly lead to labor hoarding, with a consequent fall in average productivity, and partly to an increase in unemployment, with some of the laid-off workers being put on recall and voluntarily waiting to be recalled rather than actively searching for jobs.

Conversely, the implicit agreement between firms and workers also means that workers accommodate increases in demand for the firm's output with increased effort, even in the absence of wage increases. Hence, output fluctuates more than employment over the business cycle, and the fact that the economy is in its long-run full-employment state is not a barrier to short-run increases in output.

Price stickiness[21]

NeoKeynesian theory assumes that while some goods in the economy are homogeneous and are traded in perfectly competitive markets, most goods, especially at the retail level, are differentiated by firms in some characteristic or other. Such differentiation is often in the form of differences in color, packaging, location, associated services, or just established brand loyalty. Such differentiation in practice is usually not enough to create a monopoly for the firm but enables it to function in a monopolistically competitive manner. Profit maximization by a monopolistically competitive firm implies that it is not a price taker, as are firms under perfect competition, but a price setter with a downward-sloping demand curve for its product. Consequently, increases in the price it sets do not reduce its sales to zero, nor doreductions in it allow it to capture the whole market for its industry. As a price setter, the firm sets the profit-maximizing price and supplies the output demanded at this price.

Changing the set price imposes a variety of costs, collectively known as *menu* costs. Examples of these are: reprinting price lists and catalogues, informing customers, re-marking the merchandise, etc. These costs, though often relatively small as a percentage of the price of the firm's product, can still be greater than the gain in revenue from a small price change. Further, even if there is a net gain from changing the price following an increase in demand, it may not be enough to persuade the firm to immediately raise its price, since the inconvenience and costs to the firm's customers of frequent price changes are likely to be resented. Consequently, the firm may not find it optimal to respond to demand changes with price changes unless the demand changes imply large enough price changes. Over time, as demand increases occur, the optimal price change becomes large enough for the firm to be willing to incur the menu costs and change the actual price of its product.

These arguments imply that a monopolistically competitive firm will change its price infrequently, but will respond to intervening changes in demand by changing its output at the

[21] For an exposition of this approach, see Ball *et al*. (1988).

existing price. In the long run, the price will adjust to demand and, even in the short run, if the demand increase is large enough, the price adjustment will occur. In the economy as a whole, an increase in aggregate demand will cause some sectors and firms, especially those with more competitive markets, to adjust their prices faster, while others will not immediately do so but will respond to demand changes with supply changes. Consequently, an increase in the aggregate demand will be partly met with an increase in prices and partly by an increase in output.

The increase in aggregate output to meet an increase in aggregate demand requires an increase in employment. Even if the economy was initially in its long-run state, the efficiency wage theory, unlike the classical theory, implies that there would exist, in this long-run state, the involuntary unemployment of workers and that these workers are willing to accept jobs at the existing real wage. Hence, the increase in aggregate demand will be accommodated through an increase in employment and output, without necessarily a change in real wages.

Conversely, a decrease in the demand for the products of the firm in monopolistic competition need not immediately cause it to lower its price unless the implied optimal price reduction was sufficiently large. Again, taking the economy to be a mix of firms in perfect and monopolistic competition, decreases in aggregate demand would partly result in a fall in the price level and partly in a reduction in the output supplied. The latter will cause firms to reduce their employment. However, as the efficiency wage theory argued, this fall in employment need not lead to a competitive reduction in the real wage.

Under the sticky price hypothesis, macroeconomic fluctuations, sometimes large ones, can arise from even small menu costs. Suppose that, because of a reduction in the money supply, aggregate demand falls and there is a corresponding fall in each firm's demand. If a given firm lowers its price, it moves along the new demand curve and not the old one. Its gain in revenue from this movement is relatively small and, because of menu costs, a lower price may not increase its profits, so that it does not reduce its price. Instead, it reduces its output to meet the new, lower, demand at the pre-existing price. With each firm behaving in this manner, aggregate output will fall. However, if all firms reduce their prices simultaneously, the price level will fall, real money balances in the economy will rise and the economy's and the firm's demand curves will shift back to their original position, so that there will be no drop in the firm's or the economy's output.

Sticky prices provide a justification based on menu costs for an upward-sloping short-run aggregate supply curve. This justification is different from that for the Phillips curve or its expectations-augmented version, and is also different from that for the Lucas supply rule. Its extreme version is the simplification that no firms would change prices in response to demand changes, so that the aggregate price level can be taken as constant. In this version, shown in Figure 15.4a, the price level is constant at its initial level P_0 given by the point a. The LAS curve shows the aggregate supply as y^f . Prices are sticky at P_0, so that we can specify a short-run aggregate supply curve SAS which is horizontal at P_0. An increase in aggregate demand from AD_0 to AD_1 leads to the supply of output y_1 at the sticky price P_0. Conversely, a decrease in the aggregate demand from AD_0 to AD_2 leads to the supply of output y_2, but again without an accompanying change from the sticky price level P_0. Transient and small changes in aggregate demand are therefore accommodated by the change in output, with an accompanying employment change. Cumulative changes in the same direction – or large aggregate demand changes – will, however, cause increases in prices, so that the long-run response to such changes is taken to be along the LAS curve.

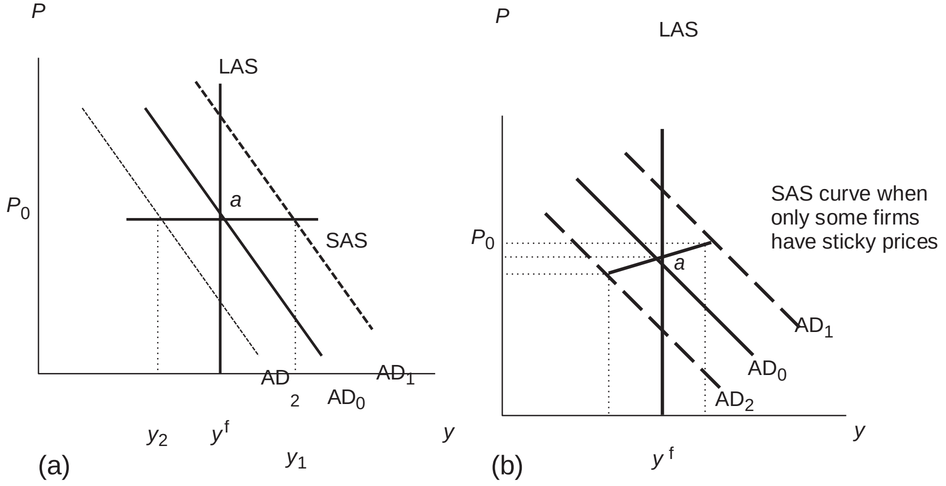


Figure 15.4

An economy with a mix of perfectly competitive (without sticky prices) and monopolistically competitive industries will have an SAS curve that is upward sloping rather than horizontal at the initial price level. Further, not all firms with sticky prices will experience sticky prices at the same time, so that the revision of prices will be staggered over time.[22] These arguments imply that the aggregate supply curve will not be horizontal but will have a positive slope, with this slope being less than if none of the firms experienced sticky prices. The nature of this curve is shown in Figure 15.4b. It assumes that outside a certain range, defined by AD_1 and AD_2, of the change in aggregate demand, it becomes profitable for all firms to change their prices. Within this range, the short-run aggregate supply curve has a positive slope with some prices remaining unchanged.

The term "menu costs" suggests small costs of changing individual prices. However, as shown above, these small costs can have "large" implications in terms of the impact of monetary policy on output and the departure from its neutrality. This result is sometimes used to make the assertion: "small nominal rigidities can have large real effects."

An implication of this menu cost approach to price stickiness is that since the firm's decision to change its price, given the menu costs, rests on whether the price change will increase or reduce its profits, it will adjust its price faster the greater the change in demand. The quicker prices adjust, the smaller will be the effect of aggregate demand increases on real output. Hence, larger changes in demand are likely to produce smaller real effects and, beyond a certain point, no increase in output. Sticky prices also reduce the real impact of demand increases in an inflationary environment. In a high inflationary environment, since prices are rising anyway, sticky price firms will find it profitable to adjust their prices more rapidly than in a zero or low inflation environment. This will reduce the increase in output.

[22] The staggered nature of this process can be captured in a Calvo adjustment process (Calvo, 1983) under which the representative firm decides on the probability (1 θ) that it will adjust its price this period and the

probability θ that it will not do so. Assuming these probabilities to be independent of the time when it last adjusted its price, the average time for which the representative firm's price remains fixed is $1/(1-\theta)$.

Importance of staggered price and wage adjustments

NeoKeynesian theories and the new Keynesian model (see below) assert that price (and nominal wage) adjustments are staggered over time rather than occurring simultaneously in a context of price-setting firms. To illustrate the importance of this point, start with an initial level of real aggregate demand that is in equilibrium with aggregate supply. Assume that an expansionary monetary policy increases aggregate demand at the existing level of prices (and nominal wages), so that relative prices (and relative and absolute real wages) are also at their initial levels. The increase in aggregate demand increases the demand for each commodity. Further, assume that the price-setting firm has increasing marginal cost. The response to the increase in aggregate demand would then be as follows.

- If all firms adjusted their prices simultaneously and continuously, all prices would rise at the same time and relative prices would not change, so that individual firms would not have an incentive to increase output. Further, the price level would rise in the same proportion as the initial increase in aggregate demand, so that aggregate demand at the now higher price level would revert to its initial real value and demand pressure on prices would be eliminated. Therefore, monetary policy would be neutral. If the initial increase in aggregate demand were due to an increase in the money supply, the real value of the money supply would revert to its initial level. If the initial increase in aggregate demand were due to a central bank's action to lower the interest rate, the price level would continue to rise until the central bank reverses its action and raises the interest rate (see Wicksell's pure credit economy analysis in Chapter 2), which is needed to return investment and consumption to their initial levels.
- If individual prices are adjusted in a staggered (rather than simultaneous) manner and discretely, rather than continuously, as in the sticky price theory, there would be slow adjustment of individual prices[23] and hence of the price level, so that the real value of aggregate demand would remain higher than in the initial equilibrium and the firms would be producing greater output. Hence, an expansionary monetary policy would have resulted in greater output during the adjustment process and would not be neutral.
- In the long run, once all adjustments are completed, all prices would have adjusted, so that relative prices and output would be as in the initial equilibrium. The price level would have increased in proportion to the increase in aggregate demand and monetary policy would become neutral.
- For a given increase in demand and, therefore, in the marginal revenue of the monopolistically competitive firm, the speed at which the price level adjusts to its long-run value depends on the elasticity of marginal cost. For a given increase in demand, the flatter the firm's marginal cost curve and the smaller the increase in the firm's price, the greater the increase in output[24] and the slower the return to long-run

[23] This process has been likened to the movement of a chain gang (whose members are tied together by a chain). Usually, the larger the number of members of the chain gang and/or the shorter the chain, the more slowly would the chain gang tend to move.

[24] This argument is sometimes stated as follows. For monopolistic firms, the profit-maximizing intersection of

the marginal cost and marginal revenue curves is below the firm’s price, so that small increases in demand will be accommodated by an output increase as long as marginal cost remains below this price. However larger demand increases that push marginal cost above the price will cause the firm to raise its price, though this is accompanied by a relatively smaller or zero output increase.

> monetary neutrality. In the context of monopolistic competition (which itself is a "small departure" from the perfectly competitive economy) and interpreting the low elasticity of marginal cost and therefore of the firm's relative price as a "real rigidity," this point is sometimes stated as the proposition: in response to changes in aggregate demand, small real rigidities can generate substantial "nominal rigidity" of the price level, which, in turn, can cause substantial departures from the neutrality of monetary policy (Blanchard, 2000, pp. 1390–91).

Note that the above conclusions pertain to monopolistic competition in commodity markets. The assumption of monopolistic competition seems to be less realistic for labor markets, and may or may not be essential to support the preceding results. However, labor markets are heterogeneous and also have departures from perfect competition and "rigidities." At least some of them are encompassed in Keynes's arguments about labor markets and the more recent implicit contract and efficiency wage theories, as well as in the theory of staggered nominal wage contracts under uncertainty (see Chapter 14 on the expectations-augmented Phillips curve).

New Keynesian (NK) macroeconomics

The new Keynesian (NK) school in macroeconomics emerged in the 1990s (Clarida, Gali and Gertler (CGG), 1999; Gali, 2002). Its general philosophy is in the Keynesian tradition and its major components are from the neoKeynesian collection of ideas. While the latter was a disparate rather than a tightly specified macroeconomic model, the NK school offers a tightly specified, integrated macroeconomic model that rebottles and re-flavors some of the neoKeynesian ideas, such as that of sticky prices, while abandoning or ignoring others, such as the efficiency wage theory and nominal wage theory. It has also adopted the Taylor rule (see Chapter 13) for the formulation of monetary policy. Its major departure from previous Keynesian economics comes from its methodology. While the methodology of earlier Keynesian thought had been somewhat eclectic and had usually used one-period comparative static analysis, the new Keynesian school adopts for its methodology stochastic, intertemporal optimization and market clearance, which are the hallmark of the modern classical school, especially its real business cycle theory. It also adopts the latter's rational expectations hypothesis (REH). For its distinctive results, the new Keynesian school relies on staggered price adjustments by monopolistically competitive firms and the assumption that the central bank sets the interest rate, not the money supply, and does it through a forward-looking Taylor rule derived from optimization of the objective function of the central bank. The resulting model has become known as the NK model. Its use of the interest rate, with the money supply made endogenous, as the critical determinant of aggregate demand, is in the tradition of Wicksell's model for the pure credit economy (see Chapter 2), so that it has sometimes been called a "neo-Wicksellian" model.

While the standard NK model, as presented below, is based on imperfections in commodity markets, it assumes that the markets for non-monetary financial assets are complete and work perfectly, with perfect substitution among all non-monetary financial assets, so that the distinction between bonds and loans/credit is irrelevant. However, another innovation by Keynesians during the last two decades has emphasized imperfect substitution between these financial assets, leading to the extension of macroeconomic modeling to incorporate credit as a separate asset from money and bonds. The distinction between these two types of assets and its implications for monetary policy are presented in the next chapter.

The following subsections present the main new Keynesian ideas on the commodity and labor markets and on monetary policy. For the formal model, it relies on CGG (Clarida *et al.*, 1999, Ch. 5). Walsh (2003) presents an in-depth review of new Keynesian models.

NK commodity market analysis

The NK closed economy (expectational) IS equation for equilibrium in the commodity market is:

$$y_t = c_t + i_t + g_t$$

where *c* is real consumption, *i* is real investment and *g* is real government expenditures on commodities. In the NK model, both households and firms pursue intertemporal optimization and use rational expectations to predict the future values of the relevant variables. Intertemporal optimization by households implies that current consumption will depend on the real interest rate and both the current and future levels of actual output (not the long-run equilibrium level), as in the life-cycle and permanent income hypotheses of consumption. Intertemporal optimization by firms implies that current investment will depend on the real interest rate and the future desired capital stock, which will depend on the future demand for commodities. Consequently, the current demand for commodities will depend on the future demand for commodities by households and firms and the real interest rate. Hence, the general form of the NK IS equation for equilibrium in the commodity market is:

$$y_t = f\,(y_{t+1},\ r_t\,,\ g_t\,)$$

where *g* is taken to represent all sources of expenditures other than consumption and investment, as well as demand shocks. The NK model of CGG states this closed-economy “expectational IS relationship” for equilibrium in the commodity market as:

$$x_t = f\,(x_{t+1},\ r_t\,,\ g_t\,)$$

where x_t y_t y^f , so that *x* represents the output gap. Assuming rational expectations and the Fisher equation, CGG approximate the preceding relationship by the log-linear expression:

$$x_t = E_t(x_{t+1}) - \psi(R_t - E_t\pi_{t+1}) + g_t \qquad (13)$$

Further, CGG assume that *g* follows a first-order autoregressive process, given by:

$$g_t = \alpha g_{t-1} + \mu_t$$

where μ is a random variable with zero mean and constant variance. Note that, since both consumption and investment are (only) forward looking and the rational expectations hypothesis is applied to their forward values, the past levels of expenditures do not affect their current levels.[25] Only the expectations of their future levels do so. Therefore, persistence

[25] Habit persistence in consumption is excluded from the model, as is the impact of past incomes through their impact on the inherited capital stock.

(i.e. the impact of the past on the present) had to be introduced into the IS equation through the specification of *g*.

Note that iterating (13) forward yields the NK IS equation as:

$$x_t = E_t\left[-W\sum_{j=0}^{\infty}(R_{t+j} - \pi_{t+1+j} + g_{t+j}\right] \tag{14}$$

NK price adjustment analysis

For the price adjustment process, the new Keynesian school assumes monopolistic competition in commodity markets and that each firm sets its product price so as to maximize its profits subject to the cost and frequency of expected future price adjustments. The latter implies that the firms set prices as in the menu-cost theory (presented earlier), with staggered price adjustments at different times by different firms in the economy.[26] However, the NK literature specifies the firm's price adjustment using a time-contingent pattern specified by Calvo (1983). Under this pattern, the representative firm decides on the probability θ that it will keep its price fixed this period, so that the probability that it will adjust its price is $(1-\theta)$.[27] θ is assumed to be independent of the time when the firm last adjusted its price, so that the adjustment made this period is independent of the past history of adjustments. Hence, the average time for which the representative firm's price remains fixed is $1/(1-\theta)$.[28] However, θ will depend on the expected future price adjustments, as will the current price adjustment.

The profit-maximizing price adjustment by a price-setting firm will depend on its marginal cost. This marginal cost rises with the level of output and the prices of inputs, which are proxied by the price level. Therefore, in logs, the representative firm's desired price p^*_t for its product can be expressed as:

$$p^*_t = P_t + \alpha x_t \tag{15}$$

where *P* is the (log of the) price level, taken to proxy input costs, and *x*, as before, is the (log of the) representative firm's share of the deviation of output from its full-employment level, with α representing the responsiveness of desired prices to the level of activity *x*. However, because price adjustment is only periodic, the price p_t to which the firm adjusts in the current period *t* is determined by a weighted average of the current and

[26] Eichenbaum and Fisher (2007) estimate, for their revision of the imperfect competition model (to allow variable elasticity of demand and firm-specific capital), that firms re-optimize prices on average every two quarters, rather than every two years, which implies a high degree of price inertia in the version without these modifications. They also find that the Calvo-style models of only periodic price adjustments can account for the behavior of post-war inflation in the USA.

[27] Alternatively, the proportion $(1-\theta)$ of firms can be assumed to change their prices in any given period, while the remainder do not do so. In this case, each firm has the same probability, $(1-\theta)$, of being a price-adjusting firm, irrespective of when it last adjusted its price. Some economists have suggested that θ should be determined endogenously and allowed to vary.

28 In a quarterly model, $\theta = 0.75$ implies that the price is adjusted on average once a year.

future expected price adjustments, so that:

$$p_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j E_t p^*_{t+j} \qquad 0 < \lambda \leq 1 \qquad (16)$$[29]

where λ is the rate at which price adjustments will be made.

The price level is the average of all prices in the economy, so that it is a weighted average of all the prices adjusted in the past. Hence:

$$P_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j E_t p_{t-j} \qquad (17)$$

Rewrite (16) and (17) as:

$$p_t = \lambda p^*_t + (1-\lambda)E_t p_{t+1} \qquad (18)$$

$$P_t = \lambda p_t + (1-\lambda)P_{t-1} \qquad (19)$$

From (15), (18) and (19), and using $\pi_t = P_t - P_{t-1}$, we get:

$$\pi_t = E_t\pi_{t+1} + [\alpha\lambda^2/(1-\lambda)]x_t \qquad (20)$$[30]

Hence, current inflation is a function of the current output gap and next period's expected inflation rate. Note that λ is the frequency of price adjustments and α is the responsiveness of the firm's desired price to the level of output relative to its full-employment level. Lower values of λ and α mean less responsiveness of inflation to current activity levels. (20) implies that higher future levels of inflation will raise the current inflation rate. Further, the lower the production relative to its full employment level, the lower will be the current inflation.

CGG state the economy's short-run supply or price/inflation adjustment equation as:

$$\pi_t = \gamma x_t + \beta E_t\pi_{t+1} + z_t \qquad (21)$$

where x is now the deviation of the economy's output from the full-employment level and z_t represents shocks, such as to the monopoly markup, that affect marginal cost; z_t is sometimes referred to as the "cost-push" element of inflation. γ is a decreasing function of θ, so that the longer the price remains sticky or unchanged, the less the elasticity of inflation to the output gap. CGG specify z_t by:

$$z_t = \rho z_{t-1} + v_t$$

[29] Desired prices further in the future have a lower weight because the possibility of price adjustments before that date is greater.

30 The price adjustment process can also be derived from the following intuitive argument. π_t is a function of π^e_{t+1} and mc_t, where *mc* is marginal cost, which affects price-adjustments since profit-maximizing firms want to equate marginal revenue to marginal cost. Future inflation affects current inflation because firms smoothen

price changes to reduce the costs of changing prices. Since marginal cost *mc* rises with an increase in output, it depends on the output gap *x*. Replacing *mc* by *x*, we get $\pi_t = f(\pi^e_{t+1}, x_t)$.

where v is a random variable with zero mean and constant variance,[31] so that z_t follows a stochastic first-order regressive process.

In terms of the impact of monetary policy on inflation, prices adjust gradually to the price level that would have occurred if the economy had remained at full employment. If firms expect a lower money supply in the future, they will also expect lower prices in the future. Their optimal response is to start by lowering current prices, which increases current real-money supply and output. Hence, the forward-looking response to future money supply shocks is to smoothen inflation over time but to make current output a negative function of future money supply changes.

New Keynesian economics considers the price adjustment equation (21) as its version of the Phillips curve, since it relates the inflation rate to the output gap, so that it is often referred to as the new Keynesian Phillips curve (NKPC). However, the original Phillips curve reflected labor market behavior, whereas the usual NK price adjustment analysis does not incorporate an explicit analysis of the labor market, though one can be appended to the NK model. This represents a glaring omission of the NK model, which is especially striking given the role assigned to labor markets and their imperfect functioning in Keynes's own analysis and those of the Keynesian models preceding the NK model. Since the analytical basis and form of (20) are derived from the optimal response to changes in demand along the marginal cost curve of the firm, rather than from the behavior of labor markets, calling it a Phillips curve is a misnomer.[32] The CGG version of the NK model does not explicitly specify labor demand and supply, and does not lay out the process for the determination of or changes in nominal or real wages. Given this, the only explicit basis for price adjustment and inflation comes from the positive slope of the marginal cost curve and the gradual adjustment of prices. Variations in work effort, as in the implicit contract and labor hoarding theory, which would produce shifts of the marginal cost curve over the business cycle, are also neglected.[33]

Further, note that given the usual upward slope of the marginal cost curve, output changes cannot occur unless they are accompanied by price changes. Therefore, (20) cannot explain how monetary policy can, at least sometimes, change output without prior or accompanying changes in prices/inflation.

The price/inflation adjustment process (20) replaces the short-run aggregate supply function of the AD–AS models, based on the labor market and production analysis in classical models and the traditional Phillips curve in Keynesian models. Iterating (21) forward yields:

$$\pi_t = E_t\left[\sum_{j=0}^{\infty}\beta^j(\lambda x_{t+j} + z_{t+j})\right] \tag{22}$$

[31] *z* can be interpreted as representing deviations from a linear impact of the output gap on marginal cost, but it can also encompass other sources of deviations. See Gali and Gertler (1999), Clarida *et al*. (CGG) (1999, page 1667, footnote 15).

[32] Equation (20) is also different from the expectations-augmented Phillips curve (EAPC) since its right-hand side has the inflation rate expected for future periods, whereas the EAPC referred to the deviation of the current inflation rate from its own expected level.

[33] Further, since the basis for inflation lies in the positive slope of the marginal cost curve, inflation could not be explained by the output gap if marginal cost was constant with respect to the output gap. This is especially likely to be so given variations in labor effort and labor hoarding.

so that the current inflation rate depends on the current and future output gaps: firms set current prices on the basis of the expectation of future marginal costs, which vary with future demand relative to the full-employment output.

Long-run supply function in the NK model

The long-run commodity supply function for the new Keynesian model, as for the modern classical model, is based on the assumptions of perfect price and wage flexibility, absence of adjustment costs, and is derived from intertemporal optimization by firms and consumers/workers. This long-run commodity supply is at the full-employment level and is independent of aggregate demand and its determinants, so that the long-run value of the output gap, x^{LR}, is zero. In addition, in (21), with stochastic errors set at zero for the long-run analysis, $z_t = z_{t-1}$. Hence, for the long run, the price adjustment equation becomes:

$$\pi^{\mathrm{LR}}{}_t = \beta E_t \pi^{\mathrm{LR}}{}_{t+1} + v_t \tag{23}$$

which makes current inflation a function only of future expected inflation, rather than of current aggregate demand and supply.

Other reasons for sticky prices, output and employment

The NK sticky information hypothesis

To get around the invalidity of the forward-looking NKPC based on sticky prices and to introduce inflation inertia, Mankiw and Reis (2002, 2006a,b) propose a staggered "sticky information" hypothesis as a basis for a price adjustment equation. Under this hypothesis, information is costly to acquire and process, so that it is updated only periodically. Adopting the Calvo process to staggered information, a fraction λ of the firms update their information each period and adjust their price, while (1 λ) are "inattentive." The adjusting firms are again drawn randomly from all firms, so that the current price level is a weighted average of past prices, rather than of the future expected price.

34

$$P_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j P_{t-j} \qquad 0 < \lambda \leq 1 \tag{24}$$

We do not specify the Mankiw and Reis model further and leave it to the reader. Suffice it to say that the price adjustment process in (24) is backward looking and generates persistence in inflation, unlike the sticky prices process. Similar sticky information processes can also be attributed to consumers and workers. While this adjustment process is closer to the traditional Phillips curve or one with static or adaptive expectations, Mankiw and Reis argue that the sticky information hypothesis provides a preferable foundation for inflation inertia in terms of microeconomic optimization.

34 In the determination of the current price level, price levels further in the past would have relatively lower weights since they would also be incorporated in more recent price levels.

Other reasons for staggered and hesitant price and production strategies

One source of staggered price and output adjustments in response to a shift in demand or supply functions is the hypothesis that the monopolistically competitive firm faces adjustment costs of changing prices, output and employment. For any one of these three variables, the simplest such hypothesis posits a quadratic cost function for the adjustment from last period's level of the variable, whose minimization implies that it will adjust partially each period[35] (see Chapter 8 for an analysis based on adjustment costs). Aggregating the adjustment cost of all three variables, cost minimization will imply the partial adjustment each period of the firm's product price, employment and output, with the changes made becoming a function of last period's price and output, so that there will occur short-run persistence, with decay, in output, employment and price variations over time.

Another source of staggered adjustments by the firm arises from the nature of the flow of information and risk aversion by firms. Leaving aside the possibility, considered above, of information being sticky, information is not only costly to process, it arrives in a discrete, staggered manner and is usually inadequate and ambiguous and often contains contradictory signals. In terms of its signal, such information is not only incomplete and fuzzy (meaning vague) but also "dirty," by which we mean that different parts of it provide signals contrary to those coming from other parts, so that the overall message is not transparent. To illustrate, suppose that the firm relies on a number of leading indicators of the economy's aggregate demand and output levels in predicting the demand for its product, these being of the type commonly used by the central bank and economic analysts. Usually, at a given time, while incoming data on some leading indicators of real output might signal a future decline in aggregate demand and output, the data on some others would point to an increase, while the relevant data on many other indicators would still not be available. Hence, altogether, the information is not only incomplete but its overall signal is fuzzy and dirty, so that there is "fundamental uncertainty" rather than merely knowable, risk-based information.[36] In such a context, suppose that the firm forms a subjective probability distribution on the change in economic activity in the quarter ahead, with the expected mean being a negative one. Being risk averse, it reduces its own output and price not by the amount that it would do if the standard deviation of the distribution was zero – that is, it was sure of a decline and of its amount – but by a lesser amount. Further, given the fuzziness and dirtiness of the signal,[37] the firm would be averse to adopting measures that would be more costly to reverse. The decline in the risk-averse firm's production may then be achieved not by laying off a sufficient number of workers, but by reducing their effort, as explained by the implicit contract theory, and any cut in its product price may be implemented by introducing discounts and special offers, rather than by a published cut in its product price. If the information over the quarter does not change, and the expectation of a decline becomes firmer, it could follow through in the following quarter with another reduction. But if the new information indicates a movement

[35] Among other studies with such an assumption, Ireland (2001) reports support for the assumption that firms face a quadratic cost of price adjustment rather than inflation adjustment.

[36] Post-Keynesians emphasized the fundamental nature of uncertainty as arising from inadequate, internally conflicting and staggered arrival of information, whose overall assessment could shift with each new bit of information. By comparison, the modern classical, and even the new Keynesian, models assume knowable, internally consistent and adequate information.

[37] Under this nature of the available information, news and rumors can play a significant part in revisions of expectations and decisions.

of the distribution to a smaller decline in economic activity than had seemed earlier, it may decide not to undertake a further reduction. However, if the revision in information indicates a pick-up in economic activity, so that the earlier assessment is now itself assessed to be incorrect, the firm can easily reverse the reduction in its own price and production. Hence, the firm's usual risk-averse strategy amounts to its pursuing hesitant, staggered price and production strategies in small steps over time. Given the collection of firms, with different sources of data (e.g. because they are in different industries) and different assessments of the subjective probability distributions, the economy as a whole would have undergone staggered and gradual price and production adjustments.

Note that this mode of information and action is no different in nature from that of the central bank, which, even with all the data and information at its disposal in modern economies, still finds the incoming information at any given point in time to be incomplete, fuzzy and dirty. Its use of its policy instrument, whether the money supply or the interest rate, also tends to be hesitant, in small steps, and often leaves the door open for no further change, another change in the same direction or a reversal of the previous change.

To conclude, there can be many sources of stickiness of prices, output and employment. Collectively, they provide a very robust basis for a gradual adjustment pattern in these variables, as opposed to complete and instantaneous adjustments, in response to demand and supply shifts.

Interest rate determination

New Keynesian economics has adopted the currently popular assumption that the central bank uses the interest rate, rather than money supply targeting, as its main instrument of monetary policy and acts as if it decides on the interest rate through a Taylor rule.[38] With the interest rate as the primary monetary policy instrument, the central bank adjusts the monetary base to ensure equilibrium in the money and other financial markets, so that the money supply becomes an endogenous variable and is no longer relevant to the determination of aggregate demand, output, employment, the price level or inflation. It can thus be removed from the macroeconomic analysis for the determination of these variables. As argued in Chapter 13, the endogenous determination of the money supply by money demand makes the LM equation and LM curve irrelevant to macroeconomic modeling.

Different forms of the Taylor rule

As discussed in Chapter 13, the Taylor rule is a feedback rule that makes the economy's interest rate a function of the output gap and the deviation of the inflation rate from the target rate, so that the interest rate responds to deviations of output and inflation from their long-run values. There are three basic forms of this rule, as follows.

Contemporaneous Taylor rule (Taylor, 1993):

$$r^{T}{}_{t} = r_0 + \alpha x_t + \beta(\pi_t - \pi^{T}) \qquad \alpha,\ \beta > 0 \tag{25}$$

[38] This is supported, among others, by Taylor (1993), Rudebusch (1995) and Clarida *et al*. (1999, 2000). Levin *et al*. (1999, 2001) report for US data that a simple version of the inflation and output-targeting rule for the US economy does quite well and that responding to an inflation forecast, not longer than a year, performs better

than forecasts of inflation farther into the future.

Backward-looking Taylor rule:

$$r^{T}_{t} = r_0 + \alpha x_t + \beta(\pi_{t-1} - \pi^{T}) \qquad \alpha,\ \beta > 0 \tag{26}$$

Forward-looking Taylor rule:

$$r^{T}_{t} = r_0 + \alpha E_t x_{t+1} + \beta(E_t \pi_{t+1} - \pi^{T}) \qquad \alpha,\ \beta > 0 \tag{27}$$

In these rules, r_0 is often taken to represent the long-run interest rate. However, since the central bank's current decision is usually relative to the previous period's interest rate, some studies modify the Taylor rules to incorporate the last period's interest rate and, in order to introduce interest rate smoothing,[39] sometimes also the deviation $Dr_{t\ 1}$ of the last period's interest rate from its long-run equilibrium level. Doing both of these modifies the above rules to the following ones.

Contemporaneous Taylor rule with persistence/smoothing:

$$r^{T}_{t} = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha x_t + \beta(\pi_t - \pi^{T})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{28}$$

Backward-looking Taylor rule with persistence/smoothing:

$$r^{T}_{t} = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha x_t + \beta(\pi_{t-1} - \pi^{T})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{29}$$

Forward-looking Taylor rule with persistence/smoothing:

$$r^{T}_{t} = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha E_t x_{t+1} + \beta(E_t \pi_{t+1} - \pi^{T})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{30}$$

In these rules, Dr_{t_1} is the deviation of the last period's interest rate from its long-run equilibrium level. The backward-looking (forward-looking) rule can be augmented to include additional backward (forward) output and inflation gaps.

Because of lags in the impact of monetary policy, a forward-looking Taylor rule is preferable to the contemporaneous and backward-looking ones. There is almost always a time lag, estimated at about six quarters or so for many Western economies, between the change in the interest rate and its impact on inflation and the output gap, so that current monetary policy should be formulated to address future inflationary pressures. If interest rates were increased in response to current inflation, their impact would occur at a future date when inflation is likely to have become different from the current rate, so that the policy could have an undesired impact. Although it is difficult to accurately predict future inflationary pressures, it may still be preferable to use a forward-looking Taylor rule.

Note that the convexity of the Phillips curve implies that inflation will react more strongly, and output less strongly, to a positive increase in aggregate demand than they would do to a corresponding decrease in aggregate demand: in absolute terms, a given increase in demand will increase inflation more than the decrease in inflation caused by a corresponding decrease in demand. Therefore, in order for the Taylor rule to reflect the asymmetric effects of aggregate demand changes on inflation and output, it too would have to beasymmetric.

39 A simple modification of the Taylor rule to allow interest rate smoothing does so by introducing the adjustment pattern: ${r^{T}}_{t} = \rho r_{t-1} + (1-\rho) r^{T}, 0 \leq \rho \leq 1$.

The form of the Taylor rule can be specified a priori or on an empirical basis, or through optimization of the central bank's objective function. The empirical basis relies on estimations, which usually favor a backward-looking rule. However, there are also two other ways of deriving the appropriate form of the Taylor rule. One of these comes from the argument that interest rate changes usually start to impact output and inflation only after several quarters and then continue to do so for several more quarters. If the central bank wants to set the current interest rate to address future output and inflation gaps, the appropriate form of the Taylor rule would have to be forward looking but with a distributed lag pattern over several quarters, quite possibly more than eight. The second form of the appropriate version of the Taylor rule is derived by optimization of a welfare loss function. The NK model takes this route. This is presented in the next section.

New Keynesian derivation of the forward-looking Taylor rule

While the Taylor rule can be stated as an institutional datum, some new Keynesians (CGG, 1999) prefer to derive it from the central bank's objective function. For this, the objective function is taken to be an intertemporal one over current and future output gaps and inflation, and is usually specified as the negative of a quadratic "loss function," so that it has the form:

$$-\frac{1}{2}E_t\left\{\sum_{j=0}^{\infty}\beta^j\left[\gamma x_{t+j}^2+\left(\pi_{t+j}-\pi^{T}\right)^2\right]\right\} \qquad (31)^{40}$$

where x is the output gap, so that the target output level is assumed to be the full-employment level. π^{T} is the target inflation rate, β is the central bank's time discount factor[41] and γ is the weight on the output gap relative to that on inflation (or the "inflation gap," defined as the difference between the actual inflation rate and the target one). γ is a function of the preference and technology parameters. If the target inflation rate generates the actual trend, the inflation gap will also represent the deviation of current inflation from the trend.

The central bank maximizes (with respect to x and π) the above objective function (i.e. minimizes the quadratic loss function) subject to the linear constraints specified by the IS equation and the price adjustment equation imposed by the economy (CGG, 1999). Doing so for the optimal or target real interest rate, and using $E_t\,\pi_{t+1}$ to represent the rationally expected future levels of inflation, yields the target interest rate as:

$$r^{T}_{t} = \alpha + \lambda_x x_t + \lambda_\pi(E_t\pi_{t+1} - \pi^{T}) \qquad \lambda_x,\ \lambda_y > 0 \qquad (32)$$

In this derivation of the central bank's policy rule, the policy responds to the expected future inflation rate rather than to the current one, unlike that in the standard Taylor rule. It is, therefore, a "forward looking" version of the Taylor rule and can be labeled the NK Taylor rule. In the long run, since the economy functions at full employment, $x^{LR} = 0$.

[40] Note, for comparison, that Rotemberg and Woodford (1999) base the central bank's objective function on the representative individual's welfare function. In this function, the deviation of inflation from its trend (rather than the inflation rate itself) has negative utility because this deviation makes it more difficult for economic agents to plan for consumption, investment and portfolio allocations. The individual's welfare also

depends on the output gap since this gap is associated with fluctuations in employment and income.

[41] Some studies identify this discount factor as the representative household's one.

Further, in the long run, since the central bank is taken to be able to achieve its target inflation rate, $(E_t\ \pi_{\underline{t\ 1}}\ \pi^{\mathrm{T}})^{\mathrm{LR}}\ \underset{=}{0}$. Therefore, $\alpha\ \underset{\mp}{r^{\mathrm{LR}}}$. Hence, the CGG "forward-looking Taylor rule" becomes:

$$r^{\mathrm{T}}{}_t = r^{\mathrm{LR}} + \lambda_x X_t + \lambda_\pi (E_t \pi_{t+1} - \pi^{\mathrm{T}}) \qquad \lambda_x,\ \lambda_\pi > 0 \tag{33}$$

where r^{LR} is determined by the long-run supply function for commodities and the IS equation (13). Since perfect capital markets are being assumed, the long-run nominal interest rate R^{LR}, with $\pi^e = \pi^{\mathrm{LR}} = \pi^{\mathrm{T}}$, is given by the Fisher equation as:

$$R^{\mathrm{LR}} = r^{\mathrm{LR}} + \pi^{\mathrm{T}}$$

Further, CGG (1999) allow for interest rate smoothing by letting the actual real interest rate be set by the central bank as:

$$r_t = \rho r_{t-1} + (1-\rho) r^{\mathrm{T}}{}_{t-1} \tag{34}$$

where $r^{\mathrm{T}}{}_{t-1}$ is specified by lagging (33).

Monetary policy in the NK model

The CGG version of the Taylor rule implies that the central bank should raise the real interest rate if a positive shock to aggregate demand increases the inflation rate above its target level and/or output above its full-employment level. Hence, since the central bank usually manipulates the nominal rate in the financial markets, it should raise the nominal rate more than the inflation rate.

For positive permanent increases in the aggregate supply of commodities, CGG assume that such a policy would raise permanent income and increase aggregate demand to the same extent, so that the output gap will not change. Nor will there be any pressure on inflation. Therefore, monetary policy would not have to respond to permanent supply shocks. However, the required assumption for this result is that the time path of the increase in aggregate demand will match that in aggregate supply.[42] While consumption does respond to permanent income, the short-run marginal propensity to consume out of permanent income is less than unity, so that consumption will rise by less than aggregate supply. If the other components of aggregate demand do not rise by exactly enough to fill the gap between increasing aggregate supply and increasing consumption, aggregate demand will increase less than aggregate supply. In this eventuality, output demand and short-run output will fall below the full-employment levels, creating a negative output gap, and will also lower inflation below its target level. Consequently, the Taylor rule will imply that the real interest rate will have to be reduced as a short-run measure. As against this tendency, the increase in the productivity of capital accompanying the increase in full-employment output may induce firms to increase investment, which will increase aggregate demand. Further, positive shocks to output are

[42] This evokes memories of Say's law, which was the assertion that increases in full-employment output will automatically increase demand to the same extent (see Chapter 18). Keynes made the objection to such an automatic response the cornerstone of his economics, and Keynesians usually adhere to it. It should not be made, explicitly or implicitly, a part of new Keynesian economics.

sometimes accompanied by increases in consumer and business confidence, with euphoria and exuberance, which can raise stock prices, investment and consumption, thereby increasing aggregate demand. Hence, there is no guarantee that aggregate supply shifts will, or usually do, increase aggregate demand to the same extent. Therefore, the short-run policy response under the Taylor rule cannot be predicted a priori but will depend on the actual increase in aggregate demand relative to the increase in aggregate supply. Hence, monetary policy will have to respond in an appropriate manner to shifts in supply.

Although NK models rely on sticky prices, they, just like the Friedman–Lucas supply rule in the absence of sticky prices, imply that the central bank can reduce inflation without any cost in terms of output as long as expectations of inflation change at the same rate as inflation itself. The latter would require the pursuit of a credible policy, announced sufficiently in advance to allow firms to change their price adjustments. If the revision in policies from past patterns of higher inflation is not sufficiently credible or is not announced in sufficient time, disinflation will impose reductions in output.

Variations of the overall NK model

The NK model has several versions, with their common feature being that there are three equations: an IS-type equation for commodity market equilibrium, a Taylor rule for the interest rate and an aggregate supply or price-adjustment (the "NK Phillips" curve) equation. To illustrate this diversity of versions, we specify below Woodford's (2007) version of the NK model.

Woodford's specification of what he calls an "intertemporal IS relation," derived from an Euler equation of the timing of aggregate expenditures, is:

$$\ln(y_t/y^f_t) = E_t[\ln(y_{t+1}/y^f_{t+1}) - \sigma[R_t - E_t\pi_{t+1} - r^n_t] \qquad (35)$$

where y is (real) output of commodities, y^f is its long-run equilibrium (full-employment) level;[43] R is the one-period nominal interest rate on riskless assets, r is the real interest rate and r^n is its long-run equilibrium (full-employment) value; and π is the inflation rate. Future values of the variables are rationally expected ones. There is a strong intertemporal element embodied in this equation. Compared with the usual IS equation that has y on the left-hand side, the preceding equation is specified in the form of the output gap y_{t+1}/y^f_{t+1} to facilitate the solution of the model.

The aggregate supply or price/inflation adjustment equation is:

$$\pi_t - \pi^*_t = \alpha\ln(y_t/y^f_t) + \beta E_t(\pi_{t+1} - \pi^*_{t+1}) + \mu_t \qquad \alpha > 0, 0 < \beta < 1 \qquad (36)$$

where π^*_{t+1} is the perceived or expected rate of *trend* inflation at time t.[44] The disturbance term μ incorporates exogenous cost-push elements. This equation is interpreted as a log-linear approximation of the staggered price dynamics in Calvo (1983). In this equation, if firms do not reoptimize their prices in a period, they automatically increase them at the trend inflation rate π^*, so that a change in their relative price only comes about through reoptimization.

[43] The long-run equilibrium values are determined by exogenous real factors such as production technology, population and household preferences.

[44] If economic agents expect that the central bank will achieve its target for the trend inflation rate, π^*_{t+1} will equal this target trend rate.

The third equation is the monetary policy one, specified by:

$$R_t = r*_t + \pi*_t + \lambda_\pi(\pi_t - \pi*_t) + \lambda_y \ln(y_t/y^f)\ _t \tag{37}$$

where $\pi *_t$ is the central bank's inflation target for period t and $r*_t$ is the bank's perceived value of the economy's equilibrium real interest rate in period t. Both these variables are taken to be exogenous, with shifts in them reflecting shifts in attitudes within the central bank. It is further assumed that $\pi *_t$ follows a random walk with mean zero, so that:

$$\pi*_t = \pi*_{t-1} + v^{\pi}_t \tag{38}$$

The complete model and its derivation and implications can be found in Woodford (2007, pp. 3–8).

Money supply in the NK model

The NK model does not explicitly have money supply in any of its (three) core equations, even though they determine the price level and the inflation rate. It would therefore seem that its determination of these variables is independent of the money supply and its growth rate, so that the central bank could ignore monetary aggregates and the money demand function altogether. It would also seem from its price adjustment equation, derived from firm's profit maximization, that the price level and the inflation rate are determined by the actions of firms in setting relative prices. These conclusions would be erroneous for the NK model.

Monetary policy in the NK model is determined by the central bank through its pursuit of a Taylor-type interest rate rule. Chapter 13 had argued that, if the central bank wants to achieve the interest rate that it sets (under this rule or otherwise), it must ensure the appropriate money supply in the economy. It had also argued that this money supply would be the one that ensures equilibrium in the money market, so that it must equal the money demanded at the set interest rate and the economy's values of the other determinants of money demand. To illustrate, assume that the money demand function has the linear form:

$$m^d = m^d(y^d, R, FW_0) = m_y y^d + (FW_0 - m_R R) + \eta_t \tag{39}$$

where $0 < m_y \leq 1$, $0 < m_R < \infty$ and η_t is a disturbance term. The definitions of the symbols in this and the following equations are as given in Chapter 13. For perfect capital markets, the Fisher equation on interest rates is:

$$R = r + \pi^e + v_t \tag{40}$$

where v_t is a disturbance term. Under a real interest rate rule, r is set at r^T by the central bank and y^d is determined above by the AD equation. Using, for illustration, the linear commodity sector model specified in Chapter 13 for the open economy, the AD equation is:

$$y^d = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^T + g + x_{c0} - x_{c\rho}\,\rho^r\} + (1/\rho^r)\cdot\{-z_{c0} + z_{cy} t_0 - z_{c\rho}\,\rho^r\}] + \mu_t \tag{41}$$

where μ_t is a disturbance term. Substituting these values of r^{T} and y^{d} in the money demand equation, we get:

$$
\begin{aligned}
m^{\mathrm{d}} = m_y\alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}} + g + x_{c0} - x_{c\rho}\,\rho^{\mathrm{r}}\} \\
+ (1/\rho^{\mathrm{r}})\cdot\{-z_{c0} + z_{cy} t_0 - z_{c\rho}\,\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R \pi^{\mathrm{e}} + FW_0 \\
+ m_y\mu_t + \eta_t - m_R v_t \qquad (42)
\end{aligned}
$$

If we assume the NK Taylor monetary policy function, r^{T} would be replaced by this function. Since the following results do not depend on the form of this policy function, we avoid writing the more complicated equation that would incorporate this function and, instead, proceed with the preceding equation.

The money market equilibrium condition for the central bank to ensure that the financial markets establish an interest rate equal to its desired target rate, is:

$$
\begin{aligned}
M^{\mathrm{s}}/P = m_y\alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\,\rho^{\mathrm{r}}\} \\
+ (1/\rho^{\mathrm{r}})\cdot\{-z_{c0} + z_{cy} t_0 - z_{c\rho}\,\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R \pi^{\mathrm{e}} + FW_0 \\
+ m_y\mu_t + \eta_t - m_R v_t \qquad (43)
\end{aligned}
$$

For any given values of P and π^{e}, this equation determines the real money supply M/P required for money market equilibrium. Note that the money market on its own cannot change P, whose movement depends upon the aggregate demand for commodities relative to their supply. Nor can the money market change the money supply M^{s}, which depends upon the monetary base M0 controlled by the central bank, or the monetary base multiplier ($\partial M/\partial$M0), which depends on the payments system and public behavior. Therefore, there is no equilibrating mechanism in the money market that will adjust M/P to real money demand, so that unless the central bank ensures that the economy has the nominal money supply required for money market equilibrium, there is a strong potential for disequilibrium in this market. Such a disequilibrium would mean that the central bank no longer has sufficient control over the market interest rate, on the basis of which the economy determines its aggregate demand, the price level and the inflation rate, so that the central bank would also lose control over these variables.

Therefore, the equilibrium condition for the money market and the ability of the central bank to ensure the required money supply are essential adjuncts of the NK model. Further, the derivations of the values of the market interest rate, the price level, the inflation rate, etc., from the NK model are conditional on the central bank's ability to manage the money supply to achieve equilibrium in the money market.[45] This ability is itself conditional on the economy's money demand function and the stochastic terms.

[45] The European Central Bank (ECB) in recent years has espoused what it calls the two pillars of monetary policy. Of these, one is "economic analysis," which assesses the short-to-medium-term determinants of price developments arising from the interplay of the supply and demand for commodities and factors, while the second one is "monetary analysis," which assesses the medium-to-long-term outlook for inflation from the long-run link between money and prices (Woodward, 2007). However, several other banks do not seem to make explicit use of the money supply data. As a result, there is considerable debate in the literature about

whether monetary data can contribute to the appropriate formulation of monetary policy and whether it is superfluous or erroneous for the ECB to rely on monetary aggregates. Our analysis indicates a dire need for the use of data on monetary aggregates even in the context of an economy which functions along the lines of the NK model, since the central

Taylor-type money supply rules and the relevance of money supply

If the money demand function were stable and all disturbance terms in (43) were equal to zero, substitution of the Taylor rule for the interest rate in the money market equilibrium condition would yield a money supply function that is itself a type of Taylor rule. For instance, substituting a Taylor interest rate rule of the form:

$$r^{T}{}_{t} = r_0 + \alpha(y_t - y^{f}) + \beta(\pi_t - \pi^{T}) \quad \alpha, \beta > 0 \tag{44}$$

for the interest rate in the preceding linear money demand function, yields:

$$m^{d} = m_y y^{d} + \{FW_0 - m_R \pi^{e} - m_R(r_0 + \alpha(y_t - y^{f}) + \beta(\pi_t - \pi^{T}))\} \tag{45}$$

so that, for money market equilibrium, the money supply rule would be:

$$M^{s} = P[m_y y^{d} + \{FW_0 - m_R \pi^{e} - m_R(r_0 + \alpha(y_t - y^{f}) + \beta(\pi_t - \pi^{T}))\}] \tag{46}$$

This equation provides the *money supply rule* that would deliver identical values of the endogenous variables, including the interest rate, under the posited equations for the economy. It is a Taylor-type rule since it makes the money supply a function of the output and inflation "gaps."

However, if the money demand function is not stable and predictable, the preceding money supply rule will have unstable and unknown values of the parameters m_y and m_r, so that the central bank will not know the precise amount of money to supply to the economy. Since this has been the case in recent decades in economies undergoing considerable financial innovation, following a money supply rule has proved to be inferior to following an interest rate rule for controlling aggregate demand.

However, this conclusion need not make the money supply redundant for monetary policy if some of the impact of the money supply on output and inflation were independent of interest rates. For the USA, while Rudebusch and Svensson (2002) find support for the redundancy proposition, Nelson (2002) and Hafer *et al*. (2007) reject such redundancy; money has an impact on output even after controlling for the impact of interest rates on output. Among the reasons offered in Chapter 16 for such a finding is that changes in the money supply could affect the cost and amount of credit, which could have an effect on output other than through interest rates.

Hence, for the NK model, we conclude that, although monetary aggregates do not appear in the NK model's equations, the conduct of monetary policy for this model is not merely the selection and pursuit of a Taylor rule for the target interest rate, it also includes the pursuit of an appropriate policy on the relevant monetary aggregates.[46] This remains so whether the money demand function is stable and known, or not stable and known. The former case

bank needs to ensure the appropriate money supply or face a disconnect between its set interest rate and the market rate.

[46] Therefore, it is erroneous to conclude that "money plays no role in the current consensus macroeconomic model and in the conduct of monetary policy." Its erroneous nature can be easily seen if the central bank were to run an experiment that lowers the interest rate but leaves the money supply unchanged or decreases it. In such an experiment, a disconnect would appear between the central bank's target rate and the market rate – and the economy's aggregate demand would evolve on the basis of the market rate.

yields a Taylor-type money supply rule, while the latter does not yield a known target for the money supply or its growth rate, nor a known Taylor-type rule for its determination.

We also note, in passing, that whether the central bank controls the money supply or the interest rate as its policy instrument, it remains pre-eminently the policy authority that can affect the inflation rate and, therefore, is the one to be held accountable for controlling inflation. Correspondingly, its credible commitment to a low inflation target is equally relevant to the achievement of low inflation under both of its operating policy targets (Woodford, 2007).

How does instability of the money demand function affect the provision of the money supply?

The preceding conclusion that the central bank must ensure, through the appropriate money supply, that the economy's interest rate is equal to its desired value holds whether the money demand function is stable or unstable. This task is easier – but still not easy because of the disturbance terms in the money demand function (43) – if the money demand function is stable and known. It is more difficult if the money demand function is unstable and unpredictable, as has been the case in recent decades in many Western economies. Further, if the monetary base multiplier (between the monetary base and the money supply) is also unstable, the achievement of the appropriate money supply through the central bank's control over the monetary base becomes even more problematical.

The central bank can attempt to provide the appropriate money supply through open market operations, reserve requirements, etc. While direct actions of this kind will have to be the main mode of meeting the money supply requirement, Chapter 13 argued that uncertainty over the money demand function and the monetary base multiplier implies that the central bank would have to give the private sector some leeway to bring about changes in the money supply, as, for example, through commercial bank borrowing from the central bank when the market rates rise above the central bank's discount rate.

Monetary aggregates, the quantity equation and inflation

The price level in the NK model is determined by aggregate demand relative to supply, and inflation in this model results from continuous demand versus supply pressures in the commodity market, so that it appears that there is no relation between monetary aggregates and the price level or inflation. Further, monetary aggregates are not in the NK model's three equations. Therefore, it seems as if the links emphasized by the quantity theory and by monetarists between money and prices, or between the money growth rate and inflation, dissolve in the context of an economy operating along the NK lines. Such conclusions are incorrect.

The quantity equation (see Chapter 2) is an identity that must hold in the NK model, as in any other theory. This identity asserts that:

$$P^{JJ} \equiv M^{JJ} - (y^{JJ} - V^{JJ}) \qquad (47)$$

where ^{JJ} stands for the growth rate and V is the velocity of circulation of money. Hence, for given values of y^{JJ} and V^{JJ}, there must be a close relation in a monetary economy between inflation and money growth. The NK model does not – and cannot – repeal the quantity equation. What it can and does do, is to make the money stock endogenous to the interest rate target and the price level, with the latter determined by the aggregate demand and supply of commodities.[47] Therefore, the more illuminating form of the quantity equation for the NK model is:

$$M^{JJ} \equiv P^{JJ} + y^{JJ} - V^{JJ} \tag{48}$$

To conclude, the distinctiveness of the NK model does not lie in that it severs the link between the monetary aggregates and the price level or inflation. Its distinctiveness on this link arises merely from its assumption that the central bank sets an interest rate target, which offers a different route for the determination of aggregate demand than a money supply target policy would. The related aspect of its distinctiveness is that the pursuit by the central bank of the interest rate as its monetary policy instrument makes the money supply endogenous to this interest rate, output and the money demand function. However, the central bank cannot just ignore the money supply: if it is not to lose control over the market interest rate, it has to ensure that the appropriate money supply is provided to the economy.

Note that (48) provides an easier route to the money supply that the central bank needs to provide: its growth rate has to equal the sum of the inflation rate and the output growth rate less the rate of change of velocity.

NK business cycle theory

The various models of the Keynesian paradigm imply that aggregate demand changes produce changes in output, as do shifts in the production function and labor supply, so that their explanations of the business cycle allow both shifts in aggregate demand and supply factors to be a potential cause of the business cycle. Therefore, from the perspective of monetary policy, monetary policy can be potentially a cause of business cycles as well as being useful in reducing their severity and duration. In fact, most central banks do habitually manipulate monetary policy to moderate inflation and the output gap, thereby subscribing to the implication of the Keynesian paradigm that monetary policy can moderate output and employment fluctuations, as well as maintain inflation within an acceptable range. This policy practice has been confirmed by the estimation of some form of the Taylor rule for many countries over many periods.

New Keynesians base their explanations of business cycles on sticky prices or sticky inflation, as opposed to the flexible price assumption of the real business cycle theory (see Chapter 14). In the NK model, the potential sources of fluctuations in output include shifts in aggregate demand due to shifts in investment, consumption, exports, fiscal and monetary policy, etc. Conversely, output fluctuations due to aggregate demand shocks can be moderated by monetary policy. Further, the new Keynesians' use of the Taylor rule on the formulation of monetary policy embeds in monetary policy an automatic stabilizer since, under this rule, an output gap (with actual output below the full-employment level) triggers an expansionary reduction in the interest rate.

Note that once both demand and supply sources of business cycle fluctuations are allowed, their relative importance is likely to vary over different business cycles and for different countries. One stance on the explanation of business cycles is the proposition that changes

[47] Note that this mode of determination of the price level is shared by the monetarist, neoclassical and Keynesian macroeconomic models.

in technology, possibly due to the revolution in information technology (IT), have been the *dominant* cause of actual business cycle fluctuations in industrialized economies in recent decades, and that while aggregate demand fluctuations can cause business cycles, they have not been the dominant source. There does seem to be substantial econometric support for this proposition.

To provide one illustration out of the many available studies on this topic, Ireland (2001) tests a new Keynesian dynamic, stochastic general equilibrium model of the business cycle for the USA over the period 1959:1 to 1998:4. His model includes sticky prices (or inflation) and adopts the nominal interest rate as the monetary policy variable, with an additional assumption on the cost of adjusting physical capital. This study reports that the model performs better with costs of adjusting prices rather than inflation. He also finds that persistence in inflation is due more to exogenous real shocks, i.e. to preferences and technology, than to large adjustment costs.

Reduced-form equations for output and employment in the Keynesian and neoclassical approaches

The fundamental implication of the various forms of Keynesian models is that, in conditions of aggregate demand deficiency and less than full employment, aggregate output depends upon aggregate demand and therefore on the demand management policy variables of fiscal expenditures, taxation and the money supply. On the impact of monetary policy on real output, the Keynesian analyses imply that:

1 This impact will depend upon the existing demand deficiency in the economy, so that a linear relationship between real output and the money supply, with constant coefficients, is not a proper representation of the Keynesian implications.

2 Both the unanticipated *and* the anticipated values of the money supply – as also of the fiscal variables – will affect output equally, as against the modern classical assertion that only the unanticipated values do so.

Keynesian reduced-form equations

A simple linear equation for capturing the dependence of output on the policy variables is:

$$dy = \lambda_g(\mathrm{D}y)dg + \lambda_M(\mathrm{D}y)\mathrm{d}M \qquad \lambda_g \geq 0 \qquad (49)$$

where D*y* is the output gap (y^f–y), λ_g and λ_M are functional symbols, and *M* is the relevant monetary policy instrument. λ_M is non-negative if this instrument is the money supply and non-positive if it is the interest rate.[48] As shown in the preceding sections, λ_g and λ_M depend upon the existing demand deficiency in the economy[49] and cannot be taken to be constants.

[48] In the demand-deficient cases, λ_g , λ_M will be non-zero, without being constant, while in the limiting case of long-run equilibrium (i.e. full employment), λ_g , λ_M 0=.

[49] This makes them state-contingent, while the Calvo price adjustment process of the NK model makes them time- contingent since, in this process, firms adjust prices on a time schedule, though it is one which is determined by the economic environment that they face.

For a dynamic context, define dy $=$ y_t $-$ $y_{t\;1}$, dg $=$ g_t $-$ $g_{t\;1}$, dM $=$ M_t $-$ $M_{t\;1}$, so that

(49) becomes:

$$y_t = [y_{t-1} - (\Lambda_g\,(\mathrm{D}y_{t-1})g_{t-1} + \Lambda_M\,(\mathrm{D}y_{t-1})M_{t-1})] + \Lambda_g\,(\mathrm{D}y_{t-1})g_t + \Lambda_M\,(\mathrm{D}y_{t-1})M_t \quad (50)$$ In

line with (50), an equation popular for empirically testing the Keynesian model is:

$$y_t = a_0 + \Lambda_g(\mathrm{D}y_{t-1})g_t + \Lambda_M(\mathrm{D}y_{t-1})M_t + \mu_t \qquad \Lambda_g\,, \Lambda_M \geq 0 \tag{51}$$

where a_0 equals $[y_{t-1} - (\Lambda_g\, g_{t-1} + \Lambda_M\, M_{t-1})]$, μ_t is a random term, and all variables are in logs. a_0 is sometimes replaced by a term of the form αy_{t-1} to capture persistent patterns in output, and y_t and $y_{t\;1}$ are sometimes defined to be deviations in output from its full-employment level. –

Another form of the above Keynesian equation uses the deviation of unemployment from its natural rate in place of the output gap, and can be specified as:

$$y_t = a_0 + \Lambda_g(u_t - u^{\mathrm{n}}{}_t)g_t + \Lambda_M(u_t - u^{\mathrm{n}}{}_t)M_t + \mu_t \tag{52}$$

where u is the unemployment rate, u^{n} is the natural (or full-employment) rate of unemployment and Λ_g and Λ_m are functional symbols. Λ_g $(u_t\ u^{\mathrm{n}}{}_t)$ and Λ_M $(u_t\ u^{\mathrm{n}}{}_t)$ need not be linear functions. – –

Comparison of the NK and modern classical estimating equations

Note that the modern classical equation corresponding to (52) is:

$$y_t = y^{\mathrm{LR}} + b_g(g_t - g^{\mathrm{e}}{}_t) + b_M(M_t - M^{\mathrm{e}}{}_t) + \mu_t \tag{53}$$

where $b_g, b_M > 0$ and the superscript e indicates the expected value of the variable in question. In expectational equilibrium with $g_t = g^{\mathrm{e}}{}_t$ and $M_t = M^{\mathrm{e}}{}_t$, (53) becomes:

$$y_t = y^{\mathrm{LR}} + \mu_t \tag{54}$$

which states the modern classical conclusion that if there are no errors in expectations, the only deviations around the full-employment output that can occur have to be random ones.

In the modern classical model, since any impact of money supply increases and fiscal deficits on output must *first* cause an increase in individual prices or the price level, the more appropriate equations for testing the modern classical model are:

$$y_t = a_0 + \gamma_1(P - EP_t) + \gamma_P EP_t + \mu_t \tag{55}$$

$$y_t = a_0 + \gamma_1(\pi - E\pi_t) + \gamma_\pi E\pi_t + \mu_t \tag{56}$$

where EP and $E\pi$ are the rationally expected values. The equilibrium version of the modern classical model implies that the estimated values of γ_P and γ_π will be zero, so that monetary policy changes will be neutral.

Empirical validity of the new Keynesian ideas

One way to judge the empirical validity of the NK model is to compare its implications with the stylized facts listed in Chapters 1 and 14 on the relationship between money, inflation and output. For the long run, the NK model holds that there is no relationship between money or inflation and output, as does the modern classical school, so that the stylized short-run facts are really the relevant ones for judging their relative validity. For the short run, the NK model clearly explains more of the stylized facts or does so better than the modern classical model. In particular, it explains the hump-shaped pattern of the impact of monetary changes on output and the longer lag in the impact on money of inflation than on output (Nelson, 1998; Sims, 1992; Christiano *et al.*, 1999). Wong (2000) finds that long-run monetary neutrality and short-run price stickiness hold for the United States. While the NK approach implies an unchanging pattern of impulse responses over time to monetary shocks, Wong finds that the actual responses differ during different episodes and are stronger for negative shocks than for positive ones.

Rudd and Whelan (2003) test the forward-looking NK output equation, which asserts that the current inflation is positively related to the future output gap. They find that this equation does very poorly for US data: empirically, current inflation is negatively, not positively, related to the future output gap. One reason for this poor performance is that there is a high degree of persistence in inflation, which depends heavily on its own lagged values (see also Maria-Dolores and Vazquez, 2006).

The NK school relies on sticky prices and nominal wages for the dynamic impact of monetary policy on output. Christiano *et al*. (2001) find that wage stickiness rather than price stickiness seems to be the more relevant factor in explaining this dynamic impact. However, Mankiw (2001, p. C52) points out that the NK price adjustment equation based on sticky prices "is completely at odds with the facts. In particular, it cannot come even close to explaining the dynamic effects of monetary policy on inflation and unemployment." Mankiw lists three invalid implications of the NKPC. One, it implies that a fully credible disinflation causes an increase in output, since its announcement leads firms to reduce their price increases even before the money supply growth rate is reduced. This causes an increase in the real money supply, which leads to higher output and lower unemployment. This is invalid since inflation rises in booms and falls in recessions and the commonly observed result of a decrease in the money supply is disinflation accompanied by a rise in unemployment. Two, in the NK price adjustment process, individual prices adjust intermittently so that the price level adjusts slowly to shocks. However, the change in prices as measured by the inflation rate adjusts instantly, so that the model does not generate persistence in inflation. What is observed in reality is that the impact of a shock to inflation builds up gradually over several quarters, so there is considerable persistence in inflation. Three, the NKPC does not generate plausible impulse response functions in inflation and unemployment following a monetary policy shock. Empirical data shows that a monetary policy shock affects unemployment for some time after the shock, while the shock has a *delayed and gradual* effect on inflation. Mankiw concludes that a backward-looking price adjustment process, with inflation inertia, is needed to explain the observed behavior of inflation.

Many empirical studies find support for some form of the Taylor rule. However, Levin *et al.* (1999, 2001) report that their estimates from US data for five macroeconomic models show that a simple version of the inflation and output-targeting rule for the US economy does quite well. Further, a simple autoregressive model of the interest rate has sometimes performed better than the Taylor rule, as Depalo (2006) reports for Japan. Moreover, estimates of the

Taylor rule tend to show that its coefficients do shift with changes in the leadership of the central bank. This is quite plausible since these coefficients reflect central bank preferences on responses to inflation and the output gap, etc., as the NK derivation of the Taylor rule shows.

Note that, to derive their particular form of the Taylor rule, NK models assume that the true forms of the central bank's objective function and the model are known. These are rarely, if ever, known, so that the superior performance of a specific optimal Taylor rule, derived from a specific model and an objective function, relative to simple Taylor rules, cannot be taken for granted. Given the uncertainty about the future course of the economy and errors in forecasting, it may be better to just specify the central bank's objectives on inflation, output and any other targets, while leaving their actual coefficients to vary with the central bank's appraisal of the circumstances. In this case, the central bank would commit itself only to pursuing certain objectives rather than to a set instrument rule such as a specific Taylor rule. Svensson (2003) provides an excellent and detailed discussion of this issue. This conclusion is consistent with that in Chapter 12 on time consistency and credibility: the conclusion there was that, given uncertainty about the appropriate form of the economy's constraints and how they might evolve over time, it might be preferable for the central bank to commit itself to objectives in the context of intertemporal reoptimization rather than to a precise time-consistent path of policies or to a policy rule.

Conclusions

This chapter has presented three versions of the Keynesian model and a neoKeynesian model. These attest to its evolutionary nature. The future is likely to see additional contributions and versions of the general Keynesian stance that money is not neutral in the short run, while it could – and is likely to be – neutral in the long run.

We arrange the concluding comments for this chapter into various categories, as follows.

Keynes on the wage bargain and the rigidity of wages

A major point of disagreement between the classical and Keynesian ideas is on the nature of the labor market. Keynes himself considered his objections to the classical model's assumptions on the labor market as being the most fundamental departures from classical ideas. He expressed his ideas on this as follows:

> But the … more fundamental objection [to the classical model] … flows from our disputing the assumption that the general level of real wages is directly determined by the character of the wage bargain. … For there may be no method available to labor as a whole whereby it can bring the wage-goods equivalent of the general level of money-wages into conformity with the marginal disutility of the current volume of employment. There may exist no expedient by which labor as a whole can reduce its real wage to a given figure by making revised money bargains with the entrepreneurs. …
>
> Though the struggle over money-wages between individuals and groups is often believed to determine the general level of real wages, it is, in fact, concerned with a different object. Since there is imperfect mobility of labor, and wages do not tend to an exact equality of net advantage in different occupations, any individual or group of individuals, who consent to a reduction of money-wages relatively to others, will suffer a relative reduction in real wages, which is a sufficient justification for them to resist it. On the other hand it would be impracticable to resist every reduction of real wages,

> due to a change in the purchasing-power of money which affects all workers alike; and in fact reductions of real wages arising in this way are not, as a rule, resisted unless they proceed to an extreme degree.
>
> (Keynes, 1936, pp. 12–15).

On the rigidity of nominal or real wages

The Keynesian nominal wage model, which assumes that nominal wages are rigid or that the labor supply depends on nominal rather than real wages, was popular as *the* Keynesian model in the early 1940s and 1950s. However, Leijonhufvud (1967) showed that Keynes did not make such an assumption, and that, for Keynes, in conditions of excessive unemployment, reducing real wages through induced inflation was preferable to nominal wage reductions (which were resented by workers and caused industrial unrest), so that the avoidance of nominal wage declines was a policy recommendation. It is now generally accepted: Keynes did not assume that nominal wages were rigid, either as an a priori assumption or because of the nature of the nominal wage bargain between firms and workers. Further, Keynes did not assume that workers based their supply behavior on nominal rather than real wages and thereby suffered money illusion.

The distinctive nature of labor markets was a fundamental part of Keynes's ideas. One interpretation of those ideas led to the Phillips curve, as illustrated by the following quotations from James Tobin (1972), an eminent economist in the Keynesian tradition of the 1960s to the 1980s.

Tobin on wages and unemployment

> Unemployment is, in this model as in Keynes reinterpreted, a disequilibrium phenomenon. Money wages do not adjust rapidly enough to clear all labor markets every day.
>
> The overall balance of vacancies and unemployment is determined by aggregate demand, and is therefore in principle subject to control by overall monetary and fiscal policy. Higher aggregate demand means fewer excess supply markets and more excess demand markets, accordingly less unemployment and more vacancies.
>
> (Tobin, 1972, pp. 9–10).

The critical role of dynamic analysis when aggregate demand falls

The introduction of dynamic analysis did away with the Keynesians' need to assume the rigidity of prices and nominal wages. For such analysis, given a fall in aggregate demand, the central issue is the nature of the individual firm's response to a fall in the demand for its product and the nature of the response of the worker who is laid off or whose job no longer seems to be secure, in a context where the numerous markets of the economy cannot realistically be assumed to come into macroeconomic equilibrium instantly. This is a shift in the debate from comparative static to dynamic analysis. There can be numerous plausible dynamic paths corresponding to any comparative static macroeconomic model, and not all necessarily lead to full employment or do so instantly. This implies a role for Keynesian demand management policies, depending upon the state of the economy and the speed at which it is expected to redress deficient demand or involuntary unemployment.

We illustrate the central issues from this perspective by the following two quotes.

Leijonhufvud on Keynes's methodology

> Keynes's theory was dynamic. His model was static. The method of trying to analyze dynamic processes with a comparative static analysis apparatus Keynes borrowed from Marshall. … The initial response to a decline in demand is a quantity adjustment. … The strong assumption of "rigid" wages is not necessary to the explanation of such system behavior. *It is sufficient only to give up the equally strong assumption of instantaneous price adjustments*.
>
> (Leijonhufvud, 1967, pp. 401–03; italics added).

Patinkin on effective demand analysis

> Involuntary unemployment can have no meaning within the confines of static equilibrium analysis. Conversely, the essence of dynamic analysis is *involuntariness*: its domain consists only of positions off the [notional] demand or supply curves. …
>
> First, we see that involuntary unemployment can exist in a system of perfect competition and wage and price flexibility. … Second, we see that a deficiency in commodity demand can generate a decrease in labor input without requiring a prior increase in the real wage rate.
>
> (Patinkin, 1965, pp. 323–24).

> And the assumption … that, granted flexibility, these [dynamic] forces will restore the economy to a state of full employment, is an assumption that … [a full employment] equilibrium position always exists and that the economy will always converge to it. More specifically, it is an assumption that just as the "market" can solve the system of excess demand equations (of the neoclassical model), when the level of real income is held constant during the *tâtonnement,* so can it solve it when the level of real income (and hence employment) is also permitted to vary.
>
> (Patinkin, 1965, p. 328).

The points made in these quotes are now generally accepted. Keynes's analysis was not comparative static or Walrasian equilibrium analysis based on the assumption of instantaneous market-clearing *price adjustments with perfect competition* in response to excess demand or supply, but focused on the *dynamic adjustments* made by firms when these conditions did not hold. If the Walrasian market adjustments in prices towards equilibrium were slow, then, quite plausibly, a shortfall in demand would result in reductions in output and employment for periods of significant length. If the economy was beset by fresh disturbances arising frequently, such as through bouts of pessimism or optimism about the future among firms' managers and households, the disequilibrium state would be a persistent phenomenon, with varying levels of employment or output.

NeoKeynesian and new Keynesian economics

NeoKeynesian economics came into being in an attempt to rebuild the Keynesian framework after the decline of faith in the 1970s in the Keynesian models and their policy prescriptions, and the resurgence of classical economics in the 1980s and 1990s. NeoKeynesian economics was not really an integrated macroeconomic model, but rather a collection of ideas and theories. One of these was the efficiency wage hypothesis, which asserted the short-run rigidity of real wages, in contrast to that of nominal wages. Another neoKeynesian theory was the menu cost theory, which provided a new basis for the short-run rigidity of prices through its hypothesis of menu costs in monopolistic competition. Its other theories included the implicit contract theory based on long-term contracts in a context of firm-specific labor skills, implying labor hoarding and variations in work effort over the business cycle.

New Keynesians seek to provide an integrated macroeconomic framework, which has adopted as its major component the staggered slow price adjustment, due to menu costs, made by monopolistically competitive firms. It ignores or downplays the efficiency wage theory and the implicit contract theory, which had been part of neoKeynesian economics. A second component of the new Keynesian model is its adoption of a Taylor rule for the pursuit of monetary policy. However, the essential distinctiveness of the new Keynesian model from neoKeynesian economics lies in its methodology, which derives the various equations of the model from microeconomic foundations with stochastic intertemporal optimization, rational expectations and general equilibrium. This methodology assigns greater weight to the impact on current values of the future values of the variables, while assigning much less weight to the impact of the past values of the variables. In doing so, new Keynesians have discarded critical parts of the earlier Keynesian models, such as labor market behavior, deficient demand analysis and involuntary unemployment, as well as the possibility of the failure of markets to clear. The new Keynesian model is a relatively new one and its empirical validity is still very much in dispute.

Mankiw (2001) summarizes the evidence on the relationship between inflation and unemployment in the following:

> Almost all economists today agree that monetary policy influences unemployment, at least temporarily, and determines inflation, at least in the short run. … Price stickiness can explain why society faces a short-run tradeoff between inflation and unemployment.
>
> The bad news is that the dynamic relationship between inflation and unemployment remains a mystery. The so-called “new Keynesian Phillips curve” is appealing from a theoretical standpoint, but it is ultimately a failure. It is not at all consistent with the standard stylized facts about the dynamic effects of monetary policy, according to which monetary shocks have a delayed and gradual effect on inflation. We can explain these facts with traditional backward-looking models of inflation–unemployment dynamics, but these models lack any foundation in the microeconomics theories of price adjustment.
>
> (Mankiw, 2001, p. C59).

And finally, what are we to believe?

The classical and Keynesian schools represent different views of the dynamic nature of the capitalist economy. The former views it as being in full-employment equilibrium or close to it, with the dynamic forces providing a strong tendency to return to full employment after any deviation. Keynesian schools allow the possible existence of full employment but are concerned that the economy does not always, or most of the time, perform at full employment. Historically, faith in these positions has tended to vary considerably. The Great Depression of the 1930s in industrial economies destroyed faith in the classical and neoclassical belief in a self-regulating economy The fairly stable macroeconomic performance of such economies in the 1950s and 1960s, though with active Keynesian demand management policies, produced shallow and short-lived recessions and led to a slow revival first of classical economics under the rubric of the Keynesian–neoclassical synthesis. The Keynesian policy errors in the 1970s, resulting in stagflation, tended to restore faith in the general classical position. While the dominant school in the last three decades of the twentieth century seemed to be the classical one, the Keynesian doctrines, rejuvenated and reformulated, seemed to reclaim dominance toward the end of the twentieth century. Its currently popular, or rather fashionable, form is that of the NK model, which adopts the market-clearance, general equilibrium, rational expectations agenda of the modern classical school, but adds to it market imperfections and sticky prices or information.

The economics profession does not possess empirical evidence sufficient to convince all economists to accept one paradigm and discard the other, and the performance of the economy seems to suit one paradigm at one time and the other one at other times, so that many economists keep an open mind, applying their overall knowledge depending upon the state of the economy. Most economists also maintain a fair degree of skepticism.

Given the uncertainty about the true nature of the macroeconomy and disagreements among economists about the correct workhorse for the economy, the practitioners of monetary policy, the central banks, do not follow a time-invariant commitment strategy to any particular model or its implications. Rather, as Chapter 12 argued, they follow a commitment strategy only with regard to their objectives. In their policies, they follow an active agenda, as the Keynesians in general recommend, for guiding the short-run evolution of output and employment through changes in the money supply and/or interest rates, while subscribing to the classical economists' belief that monetary policy cannot affect the long-run evolution of the real variables.

The current fashion in macroeconomics is to base its foundations strictly and solely on microeconomic intertemporal stochastic general equilibrium foundations. The following quote provides a caution against this practice.

> Economists often aspire to make our discipline like physics. Just as there are today two "economicses" – micro and macro … – there are also two "physicses": quantum theory, which describes the behavior of the tiniest particles of matter, and Newtonian mechanics (as amended by the theory of relativity), which applies to larger bodies. One of the challenges that physicists face is to integrate the two. As the distinguished mathematician Roger Penrose has pointed out, however, the way to do it is clearly not simply to take the principles of quantum theory and apply them wholesale to larger bodies. Doing so leads, in Penrose's classic example, to concluding that a basketball can be in two places at once. … Simply applying to aggregate economies what we know about the behavior of rational, profit- or utility-maximizing individual agents leads to patent contradictions of the economic world in which we live.
>
> (B.M. Friedman, 2003, p. 10).

Summary of critical conclusions

- ❖ An abiding theme of the Keynesian paradigm, originating with Keynes's *The General Theory*, is the failure of the economy to attain Walrasian general equilibrium. Many of its models assert that this failure is especially symptomatic of the labor market, so that involuntary unemployment is a common occurrence in real-world economies.

- ❖ Early (1940s and 1950s) Keynesian models were based on nominal wage rigidity or price illusion by labor.
- ❖ In the 1960s and 1970s, Keynesian models were often based on the Phillips curve.
- ❖ The Keynesian effective demand model posits that the rational dynamic responses by firms and households to conditions of inadequate demand and involuntary unemployment do not always take the economy to full employment or do so within an acceptable period.
- ❖ The neoKeynesian theories rely on rational long-run behavior, resulting in implicit contracts, staggered wage contracts, sticky prices, menu costs, etc.
- ❖ The new Keynesians base their macroeconomic model on the microeconomic foundation of forward-looking, optimizing economic agents holding rational expectations and with the economy operating in general equilibrium. Their distinctiveness from the modern classical model, which has a similar methodology, lies in that they assume monopolistically competitive firms, sticky prices and the Taylor rule for the central bank's monetary policy.

Review and discussion questions[50]

1. "In response to demand shocks, short-term quantity adjustments occur earlier than price adjustment at the level of both the firm and the economy." Discuss the relevant theory behind this statement. Also, discuss its empirical validity at the macroeconomic level.
2. Discuss in the context of the effective demand and Phillips curve Keynesian models: excluding dynamic effects, an increase in the stock of money and a fall in nominal wages have essentially the same effects at a time of involuntary unemployment.
3. Discuss in the context of the neo- and new Keynesian models: excluding dynamic effects, an increase in the stock of money and a fall in nominal wages have essentially the same effects at a time of involuntary unemployment.
4. (a) Describe a simple fixed-price short-run macroeconomic model (with flexible nominal wages) and compare it with a conventional market-clearing model. Compare their predictions for the effectiveness of monetary and fiscal policies.

   (b) Describe a simple short-run macroeconomic model with flexible prices but fixed nominal wages and compare it with a conventional market-clearing model (with flexible nominal wages). Compare their predictions for the effectiveness of monetary and fiscal policies.

   (c) Describe a simple short-run macroeconomic model with a fixed price and fixed nominal wages and compare it with a conventional market-clearing model. Compare their predictions for the effectiveness of monetary and fiscal policies.
5. You are given the following fixed-price, closed-economy, IS–LM model:

$$IS: \quad y = c[(1 - t_1)(y + b/r), M + b/r] + i(r, y) + g$$

$$LM: \; M = m^{\mathrm{d}}(R, y, M + b/r)$$

Fisher equation: with an exogenously specified expected inflation rate at zero.

[50] Unless a symbol is specifically defined in a question, its definition is as in the text.

The government's budget constraint is:

$$dM + db/r = g - t_1(y + b/r)$$

where b is the number of consols, each paying $1 per period in perpetuity. P is normalized to unity. Wealth is held only in money and bonds.

(a) Explain the differences between the IS and LM relationships in this question and those used in this chapter and Chapter 13.
(b) Explain the government budget constraint.
(c) Using IS–LM diagrams, derive the short-run and long-run equilibrium effects on output of a permanent increase in g financed by (i) money creation, (ii) bond creation. Under what conditions are these policies stable?
(d) How are your results affected if bonds are not part of net wealth?
(e) Does this model explain some of the differences between the Monetarists and the Keynesians on the relative efficacy of monetary versus fiscal policies?

6. Suppose that business pessimism reduces investment such that aggregate demand becomes less than full employment income at all *non-negative* rates of interest. Use IS–LM analysis to answer the following:

   (a) Are there positive equilibrium levels of y, r and P in the neoclassical model?
   (b) Are there positive equilibrium levels of y, r and P in the Keynesian fixed-price model? In the Keynesian nominal wage model (without fixed prices)?

   What processes will take the economy to these levels?
7. "From the time of Say and Ricardo the classical economists have taught us that the supply creates its own demand … (and) that an individual act of abstaining from consumption necessarily leads to … the commodities thus released … to be invested … so that an act of individual saving inevitably leads to a parallel act of investment … Those who think this way are deceived. They are fallaciously supposing that there is a nexus which unites decisions to abstain from consumption with decisions to provide for future consumption, whereas the motives which determine the latter are not linked with the motives which determine the former." (Keynes, 1936, pp. 18–21). Explain this statement.

   If investment and saving depend on different determinants, what are these determinants and how is the equality of saving and investment in the economy ensured? Or does it also become an identity? If it is not an identity, outline the possible scenario of the likely adjustment pattern in the economy following an exogenous decrease in consumption.
8. Suppose the central bank pegs the price level by using money supply changes through open market operations. Present the IS–LM analysis incorporating this money supply rule and show the implications for the money supply, aggregate demand and output of an exogenous increase in autonomous consumption. Is the effect on interest rates less or greater under this money supply rule than if the money supply were held constant.
9. Start with the neoclassical model and assume that its equilibrium solution is (y^f, n^f, r^*, P^*). Suppose a reduction in investment reduces aggregate demand. Discuss the following:

   (a) Within the context of the *neoclassical* model, analyze the behavior of firms if they face imperfect competition and are hit with a fall in the demand for their products. If this analysis shows that employment is reduced below n^f , present the likely consumption response of households. If these responses of firms and households imply a movement away from (y^f, n^f, r^*, P^*), what equilibrating mechanisms will come into play to bring the economy back to (y^f, n^f, r^*, P^*)? Which do you think is more powerful and has a faster response: the economy's equilibrating mechanisms or the (contrary) responses of firms and households which worked to take the economy away from (y^f, n^f, r^*, P^*)?

(b) In the context of a *Keynesian* model with nominal wage contracting, redo the questions in (a).

(c) Within the context of a neoKeynesian model, redo the questions in (a).

10. "The *classical theory* dominates the economic thought, both practical and theoretical, of the governing and academic classes of this generation, as it has for a hundred years past. ... [But it is] applicable to a special case only and not to a general case, the situation it assumes being a limiting point of the possible positions of equilibrium." (Keynes, 1936, p. 3).

    (a) What are the traditional classical (also neoclassical and modern classical) and Keynesian definitions of equilibrium? How are they related? Can there exist an underemployment equilibrium in their models under each of these definitions?

    (b) If you adopt the Keynesian definition of equilibrium, was the traditional classical model a special case of any of the Keynesian models? Of the new Keynesian model?

    (c) Are the neoclassical and modern classical models also special cases of any of the Keynesian models?

11. Keynes argued that an economy could be in equilibrium with a substantial amount of involuntary unemployment, but other economists took the stand that an equilibrium in which an important market does not clear is a contradiction in terms. Explain the notions of equilibrium involved, Keynes's justification for his position, and his opponents' justification for theirs.
12. Distinguish between Keynesian unemployment caused by an aggregate demand deficiency and classical unemployment due to real wages being above the full-employment level. What can monetary policy do to reduce each of these?
13. One way of capturing the degree of indexation of nominal wages is by specifying the wage contract as:

$$W - W_0 = \alpha(P - P_0) \quad 0 < \alpha < 1$$

where W_0 and P_0 are the nominal wage and price levels at the time of the negotiation of the wage contract and all variables are in logs. α 1 indicates full indexation.

Compare the aggregate supply curve when W is indexed to P with $\alpha < 1$ with the curves in the fixed nominal wage Keynesian model and the flexible nominal wage neoclassical model. What are the implications of $\alpha = 1$ for the responsiveness of output and the price level to (a) aggregate demand shocks, (b) aggregate supply shocks? In this case, would the aggregate supply curve differ from the aggregate supply curve with a flexible nominal wage?

Show that real output is less sensitive and the price level more sensitive to changes in the money supply if α is greater.

14. J. R. Hicks, in the 1937 article in which he proposed the IS–LM analysis, argued that

Keynes’s *General Theory* did not represent a major break with the classical tradition.

In particular, he maintained that the main insight it contained was into the conditions existing during a depression or a deep recession. Was this claim valid?

Have Keynesians contributed anything further since Keynes's *General Theory*? Does the above claim apply to the various Keynesian models?

Does the modern classical approach provide an adequate analysis of the economic conditions in recessions and depressions, or does the profession still need the Keynesian approaches for its analysis?

15. "Keynes argued that wage stickiness was probably a good thing, that wage and price flexibility could easily be destructive of real economic stability. His reasoning went like this. In a monetary economy, the nominal interest rate cannot be negative. Hence the real interest rate must be at least equal to the rate of deflation. … If wages and prices were to fall freely after a contractionary shock, the real interest rate would become very large at just the wrong time, with adverse effects on investment. The induced secondary contraction would only worsen the situation." (Solow, 1980). Discuss the validity of these arguments. Do they apply also to the modern classical model?
16. Keynes (1936) argued that, from a policy perspective, everything that can be achieved by a nominal wage cut can be more effectively achieved through an appropriate monetary policy.

    (a) Does this statement hold in the deficient-demand Keynesian model for a negative shock to (i) aggregate demand and (ii) aggregate labor productivity?

    (b) Does this statement hold in the new Keynesian model for a negative shock to (i) aggregate demand and (ii) aggregate labor productivity?

17. "Keynesianism and new Keynesianism are fundamentally inconsistent in so far as their wage hypotheses are concerned. Keynesianism asserts nominal wage rigidity, at least downwards, while the new Keynesianism does not allow either nominal or real wage rigidity." Discuss.
18. In the last two decades of the twentieth century, many economists believed that Keynesian economics give little or even wrong prescriptions for dealing with the current economic problems in the United States (or British or Canadian) economy. What justifies such comments? What are your views on this issue and how would you justify them?
19. How does the new Keynesian model differ from the earlier Keynesian deficient-demand model? How does it differ from the modern classical one? Which of the three models would explain involuntary unemployment in a recession following a fall in aggregate demand?
20. "Every major inflation has been produced by monetary expansion" (Friedman, 1968). Does this assertion hold for the NK model, whose equations do not even include a monetary aggregate as a variable?
21. "Because monetary shocks have a delayed and gradual impact on inflation, in essence we experience a credible announced disinflation every time we get a contractionary shock. Yet we don't get the boom that the (new Keynesian) model says should accompany it. This means that something is fundamentally wrong with this model." Discuss the validity of the preceding observation. In particular, is this observation valid for contractions in monetary policy that were unanticipated when the policy was first announced? What does the new Keynesian model imply on this issue? If it is

fundamentally wrong, how might the new Keynesian model be modified to fit the facts?

22. “It is now well established that a contractionary monetary shock increases unemployment before reducing inflation and that the peak impact on unemployment precedes the peak impact on inflation.” Discuss how well these observations are explained by (i) the modern classical model, (b) the new Keynesian model, and (c) any one other Keynesian model of your choice.

References

Ball, L., Mankiw, N.G. and Romer, D. “The new Keynesian economics and the output–inflation trade-off.” *Brookings Papers on Economic Activity*, 19, 1988, pp. 1–65.

Blanchard, O. “What do we know about macroeconomics that Fisher and Wicksell did not?” *Quarterly Journal of Economics*, 65, 2000, pp. 1375–409.

Boschen, J.F. and Weise, C.L. “What starts inflation: evidence from the OECD countries.” *Journal of Money, Credit and Banking*, 35, 2003, pp. 323–49.

Calvo, G. “Staggered prices in a utility maximizing framework.” *Journal of Monetary Economics*, 12, 1983, pp. 383–98.

Christiano, L.J., Eichenbaum, M. and Evans, C. “Monetary policy shocks: what have we learned and to what end?” In J. Taylor and M. Woodford, eds, *Handbook of Macroeconomics*, Vol. 1A. Amsterdam: Elsevier North-Holland, 1999, pp. 65–148.

Christiano, L.J., Eichenbaum, M. and Evans, C. “Nominal rigidities and the dynamic effects of a shock to monetary policy.” *NBER Working Paper no.* 8403, 2001.

Clarida, R., Gali, J. and Gertler, M. “The science of monetary policy: a new Keynesian perspective.”

Journal of Economic Literature, 37, 1999, pp. 1661–707.

Clarida, R., Gali, J. and Gertler, M. “Monetary policy rules and macroeconomic stability: evidence and some theory.” *Quarterly Journal of Economics*, 115, 2000, pp. 147–80.

Clower, R. “The Keynesian counter-revolution: a theoretical appraisal.” In F.H. Hahn and

F.P.R. Brechling, eds, *The Theory of Interest Rates*. London: Macmillan, 1965. Depalo, D. “Japan: the case for a Taylor rule? A simple approach.” *Pacific Economic Review*, 11, 2006,

pp. 527–46.

Eichenbaum, M. and Fisher, J.D.M. “Estimating the frequency of price re-optimization in Calvo-style models.” *Journal of Monetary Economics*, 54, 2007, pp. 2032–47.

Friedman, B.M. “The LM curve: a not-so-fond farewell.” *NBER Working Paper* no. 10123, 2003.

Friedman, M. “The role of monetary policy.” *American Economic Review*, 58, 1968, pp. 1–17. Reprinted in Milton Friedman, *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine Publishing Co. 1969, p. 106.

Gali, J. “New perspectives on monetary policy, inflation and the business cycle.” *NBER Working Paper* no. 8767, 2002.

Galí, J. and Gertler, M. “Inflation dynamics: a structural econometric analysis.” *Journal of Monetary Economics,* 44, 1999, pp. 195–222.

Hafer, R.W., Haslag, J.H. and Jones, G. “On money and output: Is money redundant?” *Journal of Monetary Economics*, 54, 2007, pp. 945–54.

Handa, J. *Monetary Economics*. London, Routledge, 2000.

Hicks, J.R. “Mr. Keynes and the classics: a suggested interpretation.” *Econometrica*, 5, 1937, pp. 147–59.

Ireland, P.N. “Sticky-price models of the business cycle: specification and stability.” *Journal of Monetary Economics*, 47, 2001, pp. 3–18.

Keynes, J.M. *A Tract on Monetary Reform*. London: Macmillan, 1923.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Leijonhufvud, A. “Keynes and the Keynesians.” *American Economic Review, Papers and Proceedings*, 57, May 1967, pp. 401–10.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press, 1968.

Levin, A., Wieland, V. and Williams, J.C. “Robustness of simple monetary policy rules under model uncertainty.” In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999, pp. 263–99.

Levin, A., Wieland, V. and Williams, J.C. “The performance of forecast-based monetary policy rules under model uncertainty.” Working Paper 2001-39, *Board of Governors of the Federal Reserve System*, 2001.

Mankiw, N.G. “The inexorable and mysterious tradeoff between inflation and unemployment.”

Economic Journal, 111, 2001, pp. C45–C61.

Mankiw, N.G. “Pervasive stickiness.” *American Economic Review*, 96, 2006a, pp. 164–9.

Mankiw, N.G. “Sticky information in general equilibrium.” *NBER Working Paper* no. 12605, 2006b.
Mankiw, N.G. and Reis, R. “Sticky information versus sticky prices: a proposal to replace the new

Keynesian Phillips curve. *Quarterly Journal of Economics*, 117, 2002, pp. 1295–328.

Maria-Dolores, R. and Vazquez, J. “How does the new Keynesian monetary model fit in the U.S. and the Eurozone? An indirect inference approach.” *Topics in Macroeconomics*, 6, 2006, article 9, pp. 1–49.
Nelson, E. “Sluggish inflation and optimising models of the business cycle.” *Journal of Monetary*

Economics, 42, 1998, pp. 302–22.

Nelson, E. “Direct effects of base money on aggregate demand: theory and evidence.” *Journal of Monetary Economics*, 49, 2002, pp. 687–708.

Okun, A. *Prices and Quantities: A Macroeconomic Analysis*. Washington DC: Brookings Institution, 1981.

Patinkin, D. *Money, Interest and Prices*. New York: Harper and Row, 1965.

Phillips, A.W. “The relation between unemployment and the rate of change of the money wage rates in the U.K. 1861–1957.” *Economica*, 25, 1958, pp. 283–99.

Rotemberg, J. and Woodford, M. “Interest rate rules in an estimated sticky price model.” In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999.

Rudd, J. and Whelan, K. “Can rational expectations sticky price models explain inflation dynamics.” at www.federalreserve.gov/pubs/feds/2003/200346, 2003.

Rudebusch, G.D. “Federal reserve interest rate targeting, rational expectations and the term structure.”

Journal of Monetary Economics, 35, 1995, pp. 245–74.

Rudebusch, G.D. and Svensson, L.E.O. “Eurosystem monetary targeting: lessons from US data.”

European Economic Review, 46, 2002, pp. 417–42.

Shapiro, C. and Stiglitz, J.E. “Equilibrium unemployment as a worker discipline device.” *American Economic Review*, 74, 1984, pp. 433–44.

Sims, C.A. “Interpreting the time series facts: the effects of monetary policy.” *European Economic Review*, 36, 1992, pp. 975–1000.

Solow, R. “On theories of unemployment.” *American Economic Review*, 70, 1980, pp. 1–11.

Svensson, L.E.O. “What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules.” *Journal of Economic Literature*, 41, 2003, pp. 426–77.

Taylor, J.B. “Discretion versus policy rules in practise.” *Carnegie-Rochester Conference Series on Public Policy*, 39, 1993, pp. 195–215.

Tobin, J. “Inflation and unemployment.” *American Economic Review*, 62, 1972, pp. 1–18.

Walsh, C. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press, 2003.

Wong, K. “Variability in the effects of monetary policy on economic policy.” *Journal of Money, Credit, and Banking*, 32, 2000, pp. 179–98.

Woodford, M. “How important is money in the conduct of monetary policy?” *NBER Working Paper* no. 13325, 2007.

Yellen, J.L. “Efficiency wage models of unemployment.” *American Economic Review*, 74, 1984, pp. 200–5.

16 Money, bonds and credit in macro modeling

For the short-run macroeconomic analysis, this chapter differentiates between two types of non- monetary financial assets: bonds and credit, of which bonds are long-term financial instruments and credit represents short-term assets. The distinctive aspect of credit relies upon adverse selection, moral hazard, and monitoring and agency costs, which provide a basis for credit rationing in quantity. Credit is treated in this chapter as the variable element of working capital in the short-run. The distinctive impact of credit on economic activity is designated the credit channel.

An important element of credit is bank loans. The distinctive impact of bank loans on economic activity occurs through the bank lending channel.

Key concepts introduced in this chapter

- Adverse selection
- Moral hazard
- Monitoring and agency costs
- Credit versus bonds
- Credit rationing
- Credit market equilibrium
- Bank loans
- Working capital
- Indirect production function

The IS–LM and IS–IRT models belong to a context that assumes perfect substitution among non-monetary financial assets, which means that all non-monetary assets, as well as their different types, are perfect substitutes, so that the distinctive elements of bonds, stocks and loans, as well as the distinctions between short- and long-term bonds, government and corporate bonds, etc., can be omitted from the analysis, and all such assets can be encompassed in a composite financial asset which is labeled "bonds." Hence, the assumption of perfect financial markets leaves the macroeconomic model with only two financial assets, money (which functions as a medium of payments) and bonds (which do not), and only two financial markets, which are the markets for money and bonds.[1] In the case where the central bank sets the money supply, the money market is specified by the LM equation (see Chapter 13). In the standard IS–LM analysis, an increase in the money supply decreases the interest rate on bonds, which increases investment and aggregate demand.

The extension of the perfect markets (i.e. perfect competition, perfect information and market efficiency) hypothesis to non-monetary financial assets implies the irrelevance of the composition of financial assets in firms' and households' portfolios and liabilities. For such a scenario, Modigliani and Miller (1958) provided the Modigliani–Miller theorem, which showed that, under perfect markets and ignoring differential tax treatments, the firm's combination of bond and equity financing was irrelevant to its output and employment, as well as to its profits which would depend on technology, inputs and consumer tastes. Further, Fama (1980) showed that whether the public holds money, bonds or stocks is irrelevant to real economic outcomes, which depend only on technology, tastes and resources. Hence, under the perfect financial markets hypothesis, financial assets and their distinctive characteristics were irrelevant for the real variables of the economy and there was no need to separate them into money, bonds, stocks, loans, etc. An intuitively unrealistic implication of these results is that credit crunches[2] and bank panics/runs[3] would have no impact on output and employment, even though such impact is often observed in recessions.[4]

As against the assumption of perfect markets, the central theme of neoKeynesian and new Keynesian economics is market imperfections.[5] Applied to financial markets, market imperfections between bonds, stocks and loans imply that they are not perfect substitutes for each other since each has quite distinctive characteristics. Further, each of these assets has many distinctive sub-categories. For example, among bonds there are short-term and long-term bonds, and there are low-risk bonds issued by some governments and high-risk ones issued by some corporations. Therefore, an extreme emphasis on market imperfections would lead to the consideration of very many different non-monetary financial assets.

However, aggregation is essential to macroeconomics, so that it needs to use the smallest number of composite goods that will adequately explain the desired macroeconomic aspects of the economy. For financial assets, the general consensus until about the 1980s was that two composite assets, money and bonds, are adequate for explaining the impact of monetary

[1] In the limiting case, the assumption of perfect markets, combined with perfectly competitive and efficient markets, implies the neutrality of both money and bonds, so that all financial assets become irrelevant to the determination of output and employment.

[2] A credit crunch is a sharp decline in the supply of credit, especially bank loans.

[3] A bank panic or run is characterized by a strong shift in the desired currency/deposit ratio ofdepositors.

[4] An illustration of this comes from explanations of recessions and depressions. In the context of the Great Depression of the 1930s, under the implications of perfect markets, money and shifts in the financial structure would have no impact on real variables, so that the Great Depression would have been solely due to real causes, as specified in the real business cycle theory. Further, models that allow money non-neutrality, but in which all non-monetary assets are identical, explain the monetary contribution to the Great Depression by citing only decreases in the money supply due to the collapse of the banking system as a cause or contributory factor, but otherwise deny credit shortages any role. The financial imperfections hypothesis uses both the decrease in the money supply and the credit restrictions and shortages among the causes of the Great Depression and its duration. There is now significant empirical evidence to support this position.

[5] Information imperfections that lead to special financial arrangements for some borrowers, such as in the form of loans rather than bonds, imply that the structure of the financial system determines the sources and uses of funds and can affect real outcomes in the economy.

policy on output and prices/inflation. However, some economists, especially new Keynesians, now believe that because of significant information imperfections a classification of financial assets into three or more assets can explain some monetary policy effects that are not well explained by the two-asset, money and bonds, classification. This chapter examines this issue and specifies a macroeconomic model with money, bonds and credit, which is defined to include loans. Its emphasis on imperfections in financial markets leads to models with a credit channel, in addition to the bond interest rate channel of the transmission of monetary policy on aggregate demand. In some models, information imperfections lead to *credit rationing* (Stiglitz and Weiss, 1981).

For intuition, it would be useful first to review the main characteristics of the non-monetary financial assets in the economy. Bonds in the real-world financial markets are marketable in both primary and secondary markets and their buyers and sellers incur brokerage costs. Marketed bonds do not need any specific collateral, other than the assets of the firm if it were to become bankrupt, nor do they need the credit-worthiness of the issuer of the bond to be established on a one-to-one basis. In terms of the returns on them, their coupon payments and maturity dates are set when they are issued and their price at their maturity date is known in advance. Bonds can be long-term or short-term. The latter, when issued by firms, are called “commercial paper.”

A useful definition of credit is short-term debt, including loans of various types, which needs to be rolled over after a short period, so that its amount can be taken to be variable during the short-run. This definition places commercial paper, along with loans and trade credit, among the components of credit. In line with this definition, this chapter defines “bonds” as marketable debt instruments other than commercial paper of firms, and all bonds issued by the government. Of the components of credit, bank loans (including lines of credit) are made directly to customers by financial institutions[6] (designated in this chapter as “banks”) and trade credit is extended by firms directly to their customers. More so than bonds, loans and trade credit depend on the direct (as opposed to market) evaluation by the lender of the credit-worthiness of the borrower, which is determined from the individual circumstances of the borrower as well as the state of the markets and the economy.

In general, borrowing through credit can be arranged at short notice and the terms of some of the credit are such that they can be repaid by the borrower or recalled by the lender at any time.[7] Often, for bank loans and trade credit, the credit interest rate is adjustable by the lender even during the duration that the loan is held. The coupon rate on short-term commercial paper gets adjusted at maturity, which occurs only a short period after its issue. As compared with this short-term variability, the coupon on medium- and long-term bonds cannot be adjusted until their maturity, which does not occur for quite some time after their issue. Although the demand for such bonds can fall, this triggers an increase in their yield in the secondary bond markets, but neither the pre-set time pattern of the coupon rate promised by the issuer nor the amount made available to the issuer usually changes until their maturity.

To create a strong analytical distinction between bonds and credit, this chapter assumes that these adjustments in the coupon rate and the amount lent through bonds can occur in the

[6] Kashyap and Stein (1994) show that banks in the USA dominate loan financing and the financing of small and medium-sized firms.

[7] Given this feature, for loans, the lender can re-assess the credit-worthiness of the borrower at any time and reduce or recall the loan or ask for additional collateral.

long-run but not in the short-run, so that the amount of funds available to firms from bond issues is given for short-run macroeconomic analysis. By comparison, our assumption is that both the amount of credit (including loans and commercial paper) made available to firms and the credit interest rate can be varied in the short-run. In terms of correspondence with the real-world financing arrangements, the preceding strong distinction provides some guidance on where different types of bonds should be slotted. Since our concern is with credit to the private sector, short-term corporate bonds[8] ("commercial paper," including "finance paper") need to be slotted in "credit," along with the loans made directly by banks to firms, since both provide short-term financing to firms and their interest cost to the issuer is frequently adjustable. Hence, short-run variability in the quantity and cost to the private sector is our basis for the distinction between credit and bonds, not marketability versus non-marketability; commercial paper is marketable whereas loans are not, but both are being slotted in credit. Therefore, our analytical concept of credit includes loans and short-term corporate bonds, but not longer-term bonds or government bonds. By implication, our analytical concept of bonds only includes medium- and long-term marketed corporate bonds and all government bonds.

Looking now at stocks, stocks/equities in the real-world financial markets do not have a maturity date and most types of stocks do not promise any coupon payment. Existing stocks can only be traded in the secondary/stock market at a price that continually fluctuates. Since their price depends on the expectation of uncertain future dividends, themselves dependent on the uncertain profitability of the firm issuing them, their yield usually has a higher degree of uncertainty than bonds, which have pre-specified coupon payments and a maturity date, and loans. However, unlike loans, stocks do share with bonds the characteristic that they do not need any collateral up front, nor do they need the credit-worthiness of the issuer to be established on a one-to-one basis.

Given these differences, one could argue that macroeconomic models need to treat money, credit, bonds and stocks as distinct financial assets in the sense that no pair has perfect substitution, so that the overall macroeconomic model needs to have three non-monetary assets and their three rates of return. However, doing so increases the analytical complexity of the macroeconomic analysis beyond the simplicity of the IS–LM or IS–IRT models, which have only one non-monetary financial asset and only one rate of return, so that convincing reasons in terms of the relative impact of shifts in the bonds, loans and stock markets on the economy have to be provided to justify the extension of the macroeconomic model to incorporate three non-monetary financial assets.

This chapter follows other studies that incorporate financial imperfections in limiting the classification of financial assets to just three assets, money, bonds and credit, with stocks still lumped in bonds, and builds a simple form of the macroeconomic model incorporating these three assets. Among several good reviews on the credit channel are Bernanke (1992–93), Kashyap and Stein (1993, 1997), Hubbard (1995), Bernanke *et al.* (1999) and Walsh (2003, Ch. 7).

Note that Walras's law allows one of the markets to be omitted from explicit analysis. Following the usual convention in the IS–LM and IS–IRT analyses, this chapter omits the explicit analysis of the bond market, so that the financial markets explicitly analyzed will be those of money and credit.

[8] Some types of such bonds (asset-backed bonds) require backing by other securities, which serve as collateral, which is also often needed for loans.

For aggregate demand in the macroeconomic model, this chapter draws on the model in Bernanke and Blinder (1988) for the supply of loanable funds.[9] For the aggregate supply of commodities, this chapter relies on the *indirect production function*, in which *working capital* becomes an input, along with labor and physical capital in production, with physical capital taken to be fixed in the short-run, as in the usual textbook AS–AD model. Working capital plays the role of an input in a monetary economy by facilitating the purchases of inputs (including labor, raw materials and intermediate goods) and allowing the revenue from sales to accrue to the firm with a lag. This chapter adopts the plausible hypothesis that working capital can vary in the short-run. Since working capital is an input in the indirect production function, a decrease in it reduces the purchases of labor and other inputs, and thereby reduces output.[10]

Working capital

Working capital includes all funds that the firm uses to facilitate its purchases of inputs and production, and sales of its output. The firm may obtain its working capital from its retained earnings, equity issues, bond issues and loans (including the use of overdrafts or lines of credit) – or save on the need for working capital by arranging trade credit. In the long-run, the firm can vary each of the sources of working capital. However, for the short-run analysis, the amount of working capital prearranged by the firm through its retained earnings and issue of bonds (including equities) is taken to be predetermined[11] and not variable. Hence, for the short-run of our model, the only component of working capital that is allowed to vary is that obtained through credit.[12]

Motivating the analysis of the credit market

The study of the money and financial markets would be irrelevant for the determination of output, employment and other real variables if money and non-monetary financial assets were neutral. While such a proposition is implicitly held by virtually all macroeconomic models for the hypothetical, analytical long-run, few macroeconomic models imply it for the short-run, especially in the context of uncertainty, adjustment costs and imperfect competition. In fact, the short-term behavior of the economy is clearly such as to convince the public, economic analysts, central bankers and governments that movements in money and credit, and not merely their unanticipated components, alter output and employment. This can be clearly seen from the very active pursuit of monetary policy by many central banks to stabilize the economy's output around its long-run growth path. It is also obvious from the declines in

[9] Bernanke and Blinder provide a commonly used macroeconomic model with money, bonds and loans as financial assets. However, their analysis focuses only on the aggregate demand for output. Kiyotaki and Moore (1997) present another loan-based model intended to explain credit cycles. Their models, along with others, are summarized in Walsh (2003, Ch. 7).

[10] The asset-backed commercial paper crisis in USA in 2007 illustrates this role extremely well. Cutbacks in the amount made available in this market meant a severe tightening of credit, with a subsequent cutback in production.

[11] Remember, bonds in our model have been defined to exclude issues of commercial paper as a way of supplementing working capital in the short-run.

[12] Kashyap *et al*. (1993) and Gertler and Gilchrist (1994) found that commercial paper issuance by large firms expanded during periods of tight credit while those by small firms declined, which represents a *flight to quality* (i.e. to lower risk loans) by the lenders.

output and rises in unemployment caused by currency, credit and exchange crises.[13] Money and credit are, therefore, definitely not neutral in the short-run.[14] This needs to be reflected by the macroeconomic models. The IS–LM and IS–IRT models, at best, do this job poorly and need to be modified to achieve more realistic implications on the impact of a shortage of credit on production. That is, the effects of credit are mainly due to variations in its total, rather than to changes in its composition.

Definitions of credit and loans

If we accept that credit should be treated as distinct from bonds, what should be its definition for macroeconomic analysis? Should it be defined as synonymous with bank loans, as in much of the literature on this topic? We choose to define it as all short-term loans to the private sector, whether marketable or not. As such, its major components are trade credit (provided by suppliers of commodities to buyers), short-term corporate bonds, and loans by banks and other lenders. Therefore, under our definition of credit, loans are just one component of credit.

Our choice of credit rather than merely loans as the distinctive non-monetary financial asset is partly due to our emphasis on credit as the variable component of working capital in the short-run and the impact of variations in it on the production of commodities. Another reason is our belief that information imperfections, discussed in the next section, which are usually behind the treatment of loans as a distinctive financial asset, really affect all financial assets, including trade capital, short-term bonds and long-term bonds, rather than merely loans. Admittedly, banks, when making loans on a personal basis, cope with market imperfections in a different way to financial markets, but our belief is that this difference is less important than that between short-run variability (for all forms of credit) versus non-variability (for bonds), which is what is needed to distinguish between short-run and long-run effects. Further, trade credit and bank loans suffer in a very similar fashion from market imperfections. In addition, an advantage of including bank loans, trade credit and commercial paper within the single category of credit is that if changes in one of them offset to some extent changes in another, only the net change in the total becomes relevant for the analysis.[15]

In any case, if one chooses, credit can be replaced by loans in this chapter, without changing the gist of its arguments.

Links between credit and economic activity and between credit and monetary policy: lessons of the 2007 subprime asset crisis for macroeconomic analysis

The 2007 crisis in the subprime asset-backed credit market (ABCM) in the USA, and its impact on housing construction and real economic activity generally in the USA and the

[13] To illustrate, in 2007, during the crisis in the subprime financial markets, speeches and reports on its impact on economic activity often carried headings such as “Heading for the rocks: will financial turmoil sink the world economy?”

[14] Their non-neutrality was starkly illustrated during the subprime mortgage crisis in the USA during 2007, when the reappraisal of risk on mortgage-backed securities led to the drying up of the demand for short-term securities by financial institutions and caused fears that the shortage of working capital would lead to cutbacks in production and a recession in the US economy, and in the world economy.

[15] Several studies report that a tightening of bank loans leads to an increase in commercial paper issued by larger firms.

world, illustrates how shifts in the availability of credit affect real economic activity and how the availability of money affects that of credit, so that neither credit nor money is neutral for real economic activity. House prices in the USA had shot up during 2001 to 2006 to such an extent that the rise was generally considered to be a bubble. This rise had been prompted by shifts in investment after the collapse of stock prices due to the crash in Internet stocks in 2001–02, and by exceptionally low interest rates during 2001 to 2006. As house prices rose, the demand for housing was further boosted by the practice of mortgage lenders to ease the terms for obtaining mortgages: some mortgages were for 100 percent of the purchase price of a house, even to customers who did not have the income to cover the projected monthly mortgage payments. This practice did not pose a serious problem as long as house prices continued to rise sufficiently fast, since the price increase provided a cushion for both buyers and lenders. However, house prices began to stabilize and then to fall in 2006. Further, interest rates began to rise. The result was an increase in mortgage defaults and a heightened perception of the riskiness of such mortgages.

Collections of the initial mortgages and other credit, such as installment credit on purchases of cars, were bundled into marketable securities, with the former serving as collateral for the latter. The securities usually had terms of 30 to 60 days, similar to that on Treasury bills, and were called asset-backed commercial paper. Since they carried a higher yield than Treasury bills but seemed to be equally liquid and safe investments, they proved attractive to many lenders as a component of their portfolios. They were bought by (production) firms, as well as by commercial and investment banks, etc. Once house prices in the USA began to fall and the risks of mortgage default rose, the riskiness and eventually illiquidity of such asset-backed commercial paper became apparent, so that the demand for them fell drastically, while the interest rates on them rose. With many issuers of such bonds unable to roll them over, the risk of default in this market rose significantly. This reduced the flow of funds for working capital to firms, affecting their production. The possibility of such a default also affected the holders of such securities, reducing their liquidity and profitability. Since many banks and other firms in the USA and many other countries held such securities, the crisis in the subprime (high-risk) market threatened to become a general financial crisis, which in turn threatened to deplete the amount of working capital for firms in production and send the USA and the world economies into a recession.

The subprime crisis could be viewed as one of the short-run liquidity, rather than long-term solvency, of most financial institutions holding the mortgage-backed securities. As part of attempts to limit the impact of the subprime crisis, central banks in the USA, Europe and other affected countries pursued an aggressive expansionary monetary policy by increasing the money supply through open market operations, cutting discount rates and openly calling for commercial banks to borrow funds as needed from the central bank. In the USA, the Fed increased the money supply very significantly, cut its discount rate by 1 percent (and openly encouraged bank borrowing at this rate), following this by a cut in the federal funds rate by ½ percent, with subsequent cuts occurring gradually. In spite of the aggressive actions taken by several central banks, the fallout of the subprime crisis for real economic activity remained uncertain for quite some time: while the projections of the real growth rates of the world economy were cut, such a reduction was nominal for the optimistic scenarios and drastic for the pessimistic ones, with the latter forecasting a severe world recession. However, it was soon clear that, by reducing the working capital of firms, the subprime crisis had the potential to cause a recession in output and that aggressive expansionary monetary policy did moderate but not eliminate this effect.

The day-by-day and week-by-week pronouncements of central banks and economic analysts during this period did also highlight the very considerable uncertainty of the extent of the need for monetary policy and the difficulty of formulating the appropriate policy in a timely fashion.

Briefly, firms habitually rely on credit to conduct their production, so that a credit crisis, by reducing the credit made available to firms, is likely to reduce their production in the short-run. If the credit markets do not provide the needed liquidity in sufficient amounts and in a timely manner, a credit crisis tends to evolve into an economic one, in which the economy goes into a recession. The standard IS–LM, AD–AS model does not provide a link from credit to production, so the subprime crisis cannot be explained by using it. This chapter tries to provide a model that links credit to the production activities of firms and also links credit to the money supply.

Distinctiveness of credit from bonds

Information imperfections in financial markets

The information that borrowers and lenders bring to their exchanges and the ability of borrowers to change the probability of default play an important role in the nature of credit contracts, the ability of credit markets to match borrowers and lenders, and the role played by the interest rate and credit rationing in the allocation of credit among borrowers. Given that there is a risk of default by a borrower, risk-averse lenders base their decision to lend on the expected return while borrowers need only base their decision to borrow on the interest rate that they have to pay. While the terms of the loan usually guarantee the latter, they do not guarantee the probability of not defaulting. Credit-rationing models imply that lenders will not raise interest rates beyond the point at which their expected return begins to decline, even though some (higher risk) borrowers are willing to pay higher interest rates (Jaffee and Russell, 1976; Stiglitz and Weiss, 1981). In this case, the credit market “equilibrium” would be characterized by an excess demand for credit, so that there would be quantity rationing (through the amount lent), in addition to price rationing through the interest rate.

“Credit rationing” occurs if, at the going interest rate, lenders supply a smaller amount than the borrowers want to obtain. In addition, among borrowers willing to pay the going interest rate, some borrowers receive credit while others do not. In fact, the latter may be willing to pay a higher interest rate but may still not get the credit. Hence, there is an unsatisfied demand for credit that is not eliminated by a rise in the interest rate. Consequently, equilibrium in the credit market is characterized by the equilibrium interest rate and credit rationing (Jaffee and Stiglitz, 1990; Stiglitz and Weiss, 1981).

Information imperfections[16] in financial markets occur because of adverse selection, moral hazard, and monitoring and agency costs – and because of the undeveloped and segmented nature of the financial sector of some economies. Under uncertainty, different borrowers can have different probabilities of repayment (versus default). If the lender can observe these

[16] Akerlof (1970) illustrated the role of information imperfections in the used car market by assuming that used car sellers know the quality of the car but buyers do not. The latter view lower prices set by sellers as indicative of poorer quality, so that lowering the price does not increase demand. Therefore, in the used car market, an excess supply of cars need not be cleared by a lower price. In fact, there may be no price at which supply equals demand.

probabilities and if they are not affected by the decision to lend, lenders can accurately rank the borrowers on the basis of their expected return and make loans to those who will yield the highest expected return. In practice, lenders cannot actually observe the borrower's probability of repayment or rely upon this probability remaining invariant. To illustrate, start with equilibrium in the credit market at the current *expected* return and assume that there is an unsatisfied demand by borrowers at this equilibrium. Given this unsatisfied demand at the equilibrium interest rate, suppose that the lenders were to increase the interest rate. At the higher rate, some borrowers, including those with less risky projects, might no longer find it profitable to borrow, while those with riskier projects may still be willing to borrow. This shift among the borrowers in the probability of repayment as the interest rate rises will, from the lender's viewpoint, represent *adverse selection*.[17] This shift in the mix of borrowers to those with lower probabilities of repayment will lower the lender's expected return, even though the credit interest rate has risen, so that, beyond a critical credit interest rate,[18] lenders will not find it to their advantage to raise the interest rate, or increase their loans, in spite of any unsatisfied loan demand. Rather, they will ration their funds among the borrowers on some basis, such as adequate and acceptable liquid collateral or their balance sheet position, other than merely the willingness to pay the going interest rate. This credit interest rate, then, is the equilibrium rate, though with rationing. Hence, borrower heterogeneity and adverse selection results in credit rationing at the equilibrium credit interest rate because lenders find it unprofitable to raise the credit interest rate, even in the face of the excess demand for loans.

Moral hazard arises when borrowers can choose among projects with different degrees of risk and lenders cannot monitor this choice. Higher credit rates of interest can entice borrowers to use the funds for riskier projects, which will reduce the lender's expected return. Just as in the case of adverse selection, moral hazard results in credit rationing at the equilibrium interest rate.

Further, under the imperfection of information available to the lender, once a loan has been arranged, the borrower may have an incentive to under-report the success of the project – and some borrowers may have an incentive to default on the loan. To offset this tendency, as well to contain the rise in risks due to adverse selection and moral hazard, lenders can resort to monitoring, on their own or through an agent, the projects financed with the loans. However, doing so involves some costs, usually labeled as *monitoring and agency costs*. These costs reduce the return, as well as the expected return, from the loan to the lender, while the cost to the borrower of the loan remains at the interest rate. Such costs would be especially applicable where the payment to the lender is positively related to the success of the project financed with the borrowed funds, so that the borrower would have an incentive to under-report the profitability of the project.[19] Agency and monitoring costs do not arise where the firm uses its internal funds but do arise when it borrows, so that they raise ("drive a wedge") the relative costs of external versus internal funds, with

[17] In the used car example in Akerlof (1970), as used car prices fall, sellers of poorer quality cars ("lemons") will be more willing to sell their cars than the sellers of better quality ones, so that there will be adverse selection by the sellers in the quality of the cars offered for sale.

[18] This critical rate is one at which the fall in the probability of repayment is more than proportional to the rise in the loan rate. Therefore, the lenders' expected return function has a local maximum at the critical loan rate.

[19] Among the contributions to the effect of adverse selection, moral hazard, monitoring and agency costs on credit markets are Jaffee and Russell, 1976; Jaffee and Stiglitz, 1990; Stiglitz and Weiss, 1981; Williamson, 1987 and Bernanke and Gertler, 1989).

this differential in costs increasing with the proportion of external to internal funds. Since recessions worsen firms' balance sheets and also reduce the availability of internal funds, they increase agency costs, thereby reducing investment, thereby worsening the recession. Hence, imperfect information in credit markets can amplify the impact of shocks to the economy.

From the borrowers' practical perspective, the impact of information imperfections and transactions costs means that either some firms are unable to raise funds in the bond market or they find it less costly to borrow from banks (henceforth, this term includes other providers of credit) rather than through the bond market, so they have to rely on loans and trade credit for all or some of their borrowing needs. In practice, while all firms rely to some extent on credit for short-term financial needs, small and medium-sized firms – as well as households who need to borrow – needing relatively small amounts do so to a much greater extent. They get barred from the bond markets because of the high transactions costs of bond issues. Firms can also get barred from the bond markets because of high monitoring costs, which require direct assessment by a lender before it will lend. Such firms have to rely on loans and trade credit for their external finance. On the lender's side of the credit market, providers of loans and trade credit have an advantage over the bond market because of their better ability to individually assess and monitor the creditworthiness of the borrower.[20] They also have a cost advantage in being able to lend relatively small amounts. In any case, they can effectively offer lower costs of credit than the bond market to some types of borrowers, especially small and riskier ones. This is also so for borrowing by consumers for purchases of durable goods.

To conclude, while information imperfections can play an important role in both credit and bonds markets, their role is stronger in credit markets since credit is short-term and needs to be renewed often. They strengthen the distinction between bonds and credit beyond merely the difference in marketability and provide a justification for the separation between bonds and credit markets for macroeconomic analysis. This distinction should be positively related to the degree of information imperfections. It would be more intense for small than for large firms. It would be also more intense, on the whole, in financially underdeveloped than in developed economies, so that production units in the former would have to rely more on internal sources of funds (often savings of owners and close relatives) and, when they do, rely more on small loans and trade credit provided under informal arrangements than on (marketed) bond issues.

Impact of monetary policy on firms' balance sheets and borrowers' creditworthiness

Monetary policy has both direct and indirect effects on firms' financial positions and therefore on their ability to obtain external funds.[21] Since firms usually have some variable-rate credit, a tight monetary policy, by raising interest rates, increases their interest payments. In addition, the rise in the interest rates may also cause a decrease in the prices of their assets and reduce the value of their collateral. These direct effects of monetary policy on the

[20] Over time, the bank lending to a firm acquires an information advantage relative to other banks, with other lenders left with an information disadvantage, so that the firm becomes dependent on a particular bank.

[21] Empirical evidence showing that internal finance is cheaper than external finance and that balance sheets matter is now quite well established (Bernanke, 1992–93).

firms' balance sheets are supplemented by an indirect effect, which occurs because a tight monetary policy reduces the demand for their products and their revenues. Therefore, both the direct and indirect effects of the tight monetary policy tend to reduce the creditworthiness of borrowers.

The preceding discussion provides two subdivisions of the way the credit markets affect the economy. These are:

- *The pure credit channel.* This affects the amount supplied of credit, for given creditworthiness of borrowers, as well as its interest cost. A component of the pure credit channel is the *bank-lending channel,* which emphasizes the special nature of bank credit and the role of banks in the economy.
- *The balance sheet channel.* This affects the demand for credit and the creditworthiness of borrowers.[22]

Market imperfections and the bank lending channel

Monetary policy affects the supply of bank loans and the banks' demand for short-term corporate paper. Banks rely mostly on deposits for virtually all of their funds. A contractionary monetary policy reduces these deposits. While banks are nowadays able to borrow directly in financial markets by issuing marketable liabilities, such as certificates of deposit and short-term bonds, these instruments are not perfect substitutes for deposits since they bear a higher interest rate than bank deposits, nor is the demand for such bank liabilities perfectly elastic.

Looking at the portfolios of "commercial banks" (using this term for a wide variety of financial institutions), the two distinct types of assets, in addition to reserves, held by banks are bonds and credit (which under our definition includes loans and commercial paper). In general, commercial banks usually do not hold stocks. A considerable proportion of the loans made by banks is often to small and medium-sized firms or consumers, who choose to take loans from banks which specialize in monitoring and enforcing contracts, because it is less expensive for them to borrow on a one-to-one basis than in the bond and stock markets. Large firms also finance a part of their working capital by short-term finance paper, which carries lower interest costs than longer-term bonds.

Therefore, the core aspects of the bank-lending channel are the lack of close substitutes for deposits on the liability side of the banks' balance sheets and the lack of close substitutes for credit (i.e. bank loans and short-term paper) for borrowers. As a result of the former, contractionary open-market operations shrink the banks' deposit base. If they try to offset this shrinkage by borrowing in the bond and equity markets, this increases the relative cost of their funds. Hence, banks' response to the shrinkage of their deposit base reduces the supply of loanable funds and raises the rates charged. On the borrowers' side, from the balance sheet mechanism explained above, the worsening of their balance sheets means that some of the borrowers will be squeezed out of the credit market while others will have to pay higher rates.

The relative amounts and significance of credit versus bonds is a relevant consideration in deciding whether a separate credit market needs to be considered explicitly for

[22] Since recessions reduce sales and worsen balance sheets, the worsening of the balance sheet will reduce the affected firm's access to credit and raise its cost, even though bond interest rates may have fallen.

macroeconomic analysis. Empirically, borrowing in the form of loans rather than bond issues is very much more important in countries with underdeveloped bond and stock markets than in the financially developed economies. Even for the latter, some economists claim that loans are significant enough in size and macroeconomic impact to be explicitly modeled apart from the bond market (Kashyap and Stein, 1993, 2000), although the usual mode of macroeconomic analysis chooses not to do so for such economies and uses the IS–LM or IS–IRT model.

Formally, the distinction between bonds and credit is not needed if either the borrowers or the lenders are indifferent between them: from a macroeconomic perspective, the separate modeling of bonds and credit is needed only if loans and other forms of credit are subject to quantity constraints[23] (i.e. rationed) on some basis other than their cost, and/or the credit rates of return/interest are different from and not perfectly correlated with the bond interest rate. The greater the slippage between the two rates or the more important the rationing constraint, the more relevant becomes the need for a separate analysis of the credit market. Loan rationing by banks and the differential in credit and bond interest rates are likely to be greater if the financial markets are fragmented, as often occurs in financially underdeveloped economies, or regulated. Examples of the latter are legal limitations on interest rates[24] and controls on amounts or proportions of bank portfolios allocated to credit to specific sectors of the economy. Many economists believe that in economies with competitive and efficient financial systems, the distinctive and significant role of the credit channel *vis-à-vis* the bond channel does not arise from the credit supply side, and therefore not from the role of banks in lending, but from the demand side, so that the borrowers' creditworthiness plays an important role. That is, the supply side of the credit channel, especially the lending one, is not significant while the balance sheet one is.[25]

Given our intent of formulating the simplest model incorporating credit as a distinctive financial asset, we stylize the preceding remarks in the following manner. Since different firms rely on credit to somewhat different degrees for their working capital, we attribute to the representative firm the average amount of funds raised through credit by all firms. This firm would, then, have a part of its funds raised through credit, with the remainder being from retained earnings or raised through bonds (including stocks), but the short-run variations in its working capital will come only from credit. Even for credit, for reasons of

[23] There is effective quantity rationing in the loan market if banks lend different amounts at any given loan rate. Such rationing does not occur in perfectly competitive markets, but the loan market, with one-to-one lending and customary monopolistic relationship established over time between a bank and its borrowers, is likely to display effective quantity rationing in the short-run.

[24] A historical illustration of this from the USA is Regulation Q, which limited the interest rate that banks could pay on their deposits. Such limits were eliminated in the 1970s. However, they remain common in many less developed economies.

[25] Ashcroft (2006) uses affiliation with a bank holding company as a proxy for financial constraints to study the behavior of banks in the USA. While small banks tend to react strongly to monetary policy changes, their behavior is counterbalanced by those of other banks, so that on average the distinct effect of monetary policy through bank loans is insignificant. Ashcroft and Campello (2007) examine loans extended to small local businesses by small subsidiary banks with the same holding company but operating in different geographical areas of the USA. Their findings show that cross-sectional variations are not due to the response of bank lending to monetary policy but depend on the creditworthiness of borrowers. These studies therefore support the balance sheet channel as against the bank lending channel for impact of monetary policy.

analytical simplification, our analysis will ignore short-run variations in trade credit. Our stylized definitions and assumptions for short-run analysis are as follows:

- "Banks" are defined as being retail banks and include moneylenders as well as other financial institutions that provide loans, usually on a one-to-one basis. For short-run analysis, banks are assumed not to hold bonds and stocks in their portfolios or, alternatively, these holdings are treated as constant in the short-run but variable in the long-run.
- Credit constitutes a positive fraction of the working capital of firms. A decrease in credit decreases the amount of working capital.
- While bonds and retained earnings of firms provide some fraction of the working capital of firms, the amount raised through them for working capital cannot be changed in the short-run, though it can be varied in the long-run.
- The working capital of firms is an argument in the indirect production function of the representative firm, with output a positive function of working capital.

Supply of commodities and the demand for credit

The representative firm engages in production and has to pay for its inputs, including labor services,[26] prior to sales. The funds used for such payments are the firm's "working capital."[27] If the firm holds inadequate funds relative to the payments to be made, it has to reduce its purchases of inputs, which reduces output. While such usage can be introduced through a cash-in-advance constraint, we choose the different route of introducing working capital in the indirect production function (see Appendix B to this chapter).[28] As Appendix B shows, working capital, in addition to the firm's use of labor and physical capital, is a determinant of output, so that the indirect production function of the representative firm can be specified as:

$$y = y(n, \overline{K}, k^w) \tag{1}$$

This production function is assumed to be twice differentiable, with the first derivatives being positive and the second negative. n is the number of workers, k^{w} is real working capital and $\underline{K}$ is the exogenously given (for short-run analysis) physical capital stock. Henceforth omitting $\underline{K}$ from the production function, this function becomes:

$$y = y(n, k^w) \tag{2}$$

Our assumptions above were that part of the working capital is raised through some combination of bonds, retained earnings and loans. Since the use of working capital arises from the need to pay labor (and for other inputs), profit maximization by the firm implies that the demand for working capital will depend on the real wage rate and on the real interest

[26] The following analysis uses the number of workers employed as the proxy for all inputs, and the wage rate as the proxy for their cost.

[27] The notion of working capital is not new in the literature. To illustrate, Keynes (1937) used loans as a constraint on investment undertaken by firms. It has also been used as a variable in the production function in some studies.

[28] This is justified by the argument that if the working capital is inadequate, the firm has to take some of its workers from production and divert them to facilitating purchases of inputs and sale of output.

rate levied on the funds raised for this purpose.[29] Hence, the demand for working capital is given by:

$$k^{w,d} = k^{w,d}(w, r^L) \qquad (3)$$

$$- \ -$$

where r^L is the interest rate on credit, which is the short-run variable part of working capital. The signs under the variables are those of the respective partial derivatives. $\partial k^{w,d}/\partial w$ is negative since an increase in the wage rate reduces the employment of labor, which reduces the need for working capital. Since r^L is the interest cost of the working capital financed through credit, $\partial k^{w,d}/\partial r^L$ is negative.

One of our stylized assumptions above was that, in the short-run, changes in the supply of or demand for bonds do not affect the amount of funds made available to a firm, nor do such changes affect the interest cost of the pre-existing bonds issued by the firm, so that, for short-run analysis, this cost becomes part of the fixed cost of the firm. Therefore, short-run production analysis only needs to take account of the real cost of credit but not of the bond rate, thereby making output, the demand for labor, real working capital and credit functions of the real wage rate and the real credit interest rate.

Supply function for commodities

For the short-run indirect production function (2), the Euler conditions for profit maximization by the firm yield the demand for labor as $n^d \ n^d(w, \ r^L)$. Assuming the simple labor supply function as $n^s \ n^s(w)$, labor market equilibrium implies that $n \ n(r^L)$. Substituting this function in the production function yields the supply of output as:=

$$y^s = y^s(r^L) \qquad (4)$$

The reason why $\partial y^s/\partial r^L \leq 0$ is that an increase in r^L reduces the working capital used by the firm, which, given the specification of the indirect production function, reduces its output. $\partial y^s/\partial r^L = 0$ if the working capital held by the firm exceeds the maximum needed for purchases of labor and other inputs. The analysis of this maximum is illustrated in Appendix A.

Demand function for credit

Further, for short-run analysis, profit maximization by the firm with the posited short-run production function implies, as shown in Appendix B, that the bond rate can also be omitted from the demand function for loans, so that:

$$L^d/P = \psi(w, R^L) \qquad (5)$$

$$- \ -$$

where L^d is the nominal demand for credit and $\psi(w, R^L)$ is the real demand. Since we do not wish to explicitly model the labor market in this chapter, we replace the wage rate by output. With an inverse relationship between w and y, rewrite (5) as:

$$L^d/P = \psi(y, R^L) \qquad (6)$$

$$+ \ -$$

[29] Appendix B expands on the nature and role of working capital in production, and its cost.

Note that, as shown in Appendix A, the firm's demand for credit has an upper limit defined by its overall need to finance the purchases of inputs. However, its actual demand for credit will be less than this limit since some of the working capital needs will have been prearranged by other sources, such as bonds and retained earnings.

Aggregate demand analysis incorporating credit as a distinctive asset[30]

Commodity market analysis

The commodity market of our open economy model is specified in the customary manner and encompassed in the standard IS equation and curve. Drawing on the familiar analysis of the IS market for the open economy, the general form of the IS equation for the open economy is:

$$y = y(\underset{-}{r}, \underset{-}{r^{\mathrm{L}}}, \underset{-}{P}) \tag{7}$$

where P is the domestic price level. In this function, the dependence of y on both r and r^{L} occurs because the real cost of raising funds for investment and working capital is given by r for the proportion raised by the representative firm through bonds (with the same rate acting as the opportunity cost of using retained earnings) but by r^{L} for the proportionraised through credit.[31] The dependence of y on P in the open-economy IS equation occurs because of substitution between domestic and foreign commodities (see Chapter 13).[32] The IS curve related to (7) has the usual negative slope whether r or r^{L} is on the vertical axis, and shifts with a change in P.

Money market analysis

Money market equilibrium

Assume that households and (production) firms (as against banks and moneylenders) do not have the choice of making loans[33] but can hold, among their assets, bonds (which include savings deposits, money market mutual funds, stocks, etc.) as the alternative to holding money. Assuming, as usual, an exogenous pre-determined level of financial wealth, FW_0,

[30] The basic structure of the commodity and money markets, and of the specification of bank loan supply, draws heavily on Bernanke and Blinder (1988). These authors point out that the standard IS–LM model incorporates an asymmetrical treatment of the banks' liabilities and assets. The former, through checkable deposits, are incorporated in the IS–LM model while the asset side is ignored. Their model modifies the IS–LM component of macroeconomic models, but not the output supply analysis of these models.

[31] If the credit market has quantity rationing in addition to that exercised by the loan rate, so that different amounts of credit are made at a given interest rate and the amount lent affects investment or consumption, the amount loaned becomes an additional argument in the IS function.

[32] This occurs when purchasing power parity (PPP) is not imposed on the model. PPP rarely holds in the economy even over long periods, and the reversion to it is remarkably slow. Its assumption is inappropriate in a short-run model.

[33] This assumption could be questionable for economies with large informal financial markets.

to be allocated to money and bonds, the standard analysis of the demand for real balances (m^d) specifies the money demand function (see Chapter 13) as:

$$m^{d} = m^{d}(\underset{+}{y}, \underset{-}{R}; \underset{+}{FW_0}) \tag{8}$$

where m^d is real money demand, y is real national income and R is the nominal bond rate. For perfect capital markets, R and r are related by the Fisher equation:

$$(1 + R) = (1 + r)(1 + \pi^{e}) \tag{9}$$

where π^e is the expected rate of inflation. Assuming π^e to be exogenously set (for simplification, at zero) or choosing to ignore it because our analysis will be a comparative static one, $(1 + \pi^e)$ is suppressed in our further analysis, so that we treat R and r as being equal (Bernanke and Blinder, 1988). Hence, the preceding money demand equation becomes:

$$M^{s} = P \cdot m^{d}(y, r; FW_0) \tag{10}$$

For a given money supply M^s, money market equilibrium specifies the usual LM equation as:

$$M^{s} = P \cdot m^{d}(y, R; FW_0) \tag{11}$$

Defining M^s narrowly as the sum of currency C in the hands of the public and bank deposits D, M^s C D. However, we also have M0 C $(RR$ $FR)$, where M0 is the monetary base M0. Therefore, the money supply is determined (see Chapter 10) as:

$$M^{s} \equiv \Sigma \frac{M0}{\frac{C}{M} + \frac{(RR+FR)}{D} - \frac{C}{M} \cdot \frac{(RR+FR)}{D}} \Sigma \tag{12}$$

where C/M is the currency ratio, RR/D is the required reserve ratio, FR/D is the free reserve ratio and $(RR+FR)/D$ is the (actual) reserve ratio. Hence, the form of the LM equation that is more appropriate for the pursuit of monetary policy than one with an exogenous money supply is:

$$\Sigma \frac{M0}{\frac{C}{M} + \frac{(RR+FR)}{D} - \frac{C}{M} \cdot \frac{(RR+FR)}{D}} \Sigma = P \cdot m^{d}(y, r) \tag{13}$$

The central bank controls the monetary base M0 and the required reserve ratio RR/D, while the public determines the currency ratio C/M and the banks determine the free reserve ratio FR/D.

Supply of bank loans

In the short-run, the supply of credit (loans, short-term commercial paper, trade credit, etc.) comes from various sources: it is provided by financial institutions, markets in short-term corporate bonds and suppliers to buyers of commodities. Of these, it is easiest to model the supply of credit by banks, so that the supply of credit is usually modeled as the supply of loanable funds by banks. Except for small other items, banks hold required reserves (*RR*), free reserves (*FR*), bonds (*B*) and credit (including loans) *L* on the asset side of their balance sheet, and have deposits *D* as liabilities, so that their balance sheet provides the identity:

$$B^{\mathrm{d}} + L^{\mathrm{s}} + FR \equiv (1 - \gamma)D$$

where γ is the required reserve ratio *RR/D* and the superscripts d and s refer to demand and supply respectively. Under our definitions above, *B* refers to longer-term bonds and *L* (for credit) includes loans and short-term commercial paper. For a given monetary base M0, an increase in the banks' desired or required reserves or in the currency/demand deposit ratio reduces the amount banks allocate to credit and bonds and raise the loan and bond rates. An increase in the monetary base increases the money supply, bank reserves, the demand for bonds and the supply of credit. The latter reduces the return on both these assets. The allocation of funds between bonds and credit depends on their relative risks, which, as discussed earlier, depend on the creditworthiness of the borrowers. An increase in the relative riskiness of credit, as caused by a recession or other adverse shocks to the balance sheets of the borrowers, will decrease the amount of credit provided by banks and increase that of bonds, representing a "flight to quality."

In view of the banks' balance sheet identity above, the impact of monetary policy on their assets depends on its impact on their budget constraint, which depends on their ability to substitute between the various types of their liabilities. Among these liabilities are demand and savings deposits, which are initiated by the public, and other short-term liabilities (such as marketed certificates of deposit (CD)) initiated by the banks. If banks can protect the sum of their liabilities from restrictive monetary policies, for example by offsetting a policy-induced fall in the monetary base and deposits by increasing the sales of their marketable liabilities domestically or by borrowing abroad, they will eliminate the impact of the monetary policy on the credit provided by banks – and, through the bank-lending channel, on the credit interest rate. This is more likely to occur in financially developed economies than in underdeveloped ones. Our assumption on this is that the banks cannot do so even in the financially developed economies, so that restrictive monetary policies reduce banks' deposits and their supply of credit, and raise the credit interest rate.

Credit market analysis

Supply of credit

Portfolio selection behavior by deposit-taking banks implies that their holdings of free reserves, credit and bonds depend on the nominal rate of return *R* on bonds, the nominal rate of return R^{L} on credit and on their "scale variable" $(1 - \gamma)D$, so that the banks' supply function for credit is:

$$L^{\mathrm{s,B}} = \lambda(\underset{+}{R^{\mathrm{L}}}, \underset{+}{R})(1 - \underset{-}{\gamma})(\underset{+}{D}) \qquad (14)$$

where $L^{\mathrm{s,B}}$ is the supply of funds by commercial banks in credit markets.

The financial system in virtually all economies has a complex, layered structure, with some financial institutions borrowing from others, and the latter borrowing from still others.

If we place commercial banks, which take deposits from the public, at the apex[34] of a triangle designating this system, we can envisage lower layers of financial institutions borrowing from those above it. The earlier analysis of market imperfections, leading to adverse selection, moral hazard, and monitoring and agency costs, applies to such borrowing by the lower layers from higher layers of financial institutions, so that the allocation of such credit among the layers is characterized by both the credit interest rate and credit rationing. It will also depend on the net worth of the borrowing units and the perceived riskiness of lending to them. These will in turn depend on the composition of their portfolios of assets. At the bottom of the triangle depicting the layers of financial institutions, the assets of such institutions will consist of credit provided to production firms and households, so that the riskiness of credit provided to financial institutions will depend on the riskiness of credit to firms and households. Hence, the supply of credit by the financial system cannot be analyzed purely from the perspective of the balance sheet of the commercial banks, so that even if the monetary base does not change, the supply of credit could vary with shifts in the riskiness of credit along the layers of the financial system.

Somewhat distinct from the layers of financial institutions is the layered structure of credit itself. In developed financial systems, credit is more than loans extended by a commercial institution that has a fair knowledge of the risks in them and holds them to maturity. The initial credit or loans can be bundled in various ways and resold as marketable securities, so that the underlying risk can become obscured, even to other financial institutions. Further, the marketable securities thus created can, in turn, be used as collateral/backing for borrowing by their holders, thus leading to a multiple creation of credit that is not captured by the bank deposit creation process, as well as to further lack of transparency of risk through the credit layers. As discussed above in this chapter, the asset-backed commercial paper crisis, originating in the USA in 1970 and spreading to many other developed countries, highlighted such multiple creation of credit, along with the potential for obscured risks along the layered structure of the credit system.

Therefore, the supply of credit is more than simply a matter of loans being extended to ultimate borrowers by the banks themselves, with the risks in such loans evaluated reasonably well by the lending bank. With risk obscured in the layers of the credit supply, shifts in the creditworthiness of borrowers can reverberate through the credit supply chain and impact on both the credit interest rate and, sometimes more so, on the quantity of credit extended at the various layers,being eventually reflected in changes in credit rationing to production firms and households. Therefore, we specify the supply of credit as:

$$L^s = L^s(\Lambda(R^L, R)(1-\gamma)D, \rho) \tag{15}$$

$$+ \quad + \qquad - \qquad +$$

where L^s is the overall supply of funds in credit markets and ρ is a (very simplistic) proxy for risk, credit rationing and structure of the credit system, so that shocks to ρ can come from shocks to any of these factors. If ρ does not change, fluctuations in credit supply will mainly reflect fluctuations in the supply of bank loans.

[34] In a system where the central bank is the lender of last resort, the central bank is really at the apex.

Demand for credit

The demand function for the nominal value L^d of credit was derived earlier from production analysis as:

$$L^d = P \cdot \psi(\underset{-}{R^L}, \underset{+}{y}) \tag{16}$$

As discussed earlier, under imperfect information, the creditworthiness of the borrower is a significant element in addition to the promise to pay the interest payments on loans and determines the expected return to the lender, as encompassed in the balance sheet channel discussed earlier. For simplicity, creditworthiness may be taken to be proxied by current output y, though it will also depend on the expected future output, so that creditworthiness is not entered as an additional argument in the credit demand function.

Credit market equilibrium

From the above, the credit market equilibrium condition is:

$$P.\psi(R, y) = L^s[\lambda(R^L, R)(1-\gamma)D, \rho] \tag{17}$$

We will designate this equilibrium equation for the credit market as the "*CC equation*."[35] Since it was assumed earlier that the expected inflation rate is exogenously given and, for convenience, set at zero, $R^L = r^L$ and $R = r$. Hence, the CC equation can be rewritten as:

$$P.\psi(r^L, y) = L^s[\lambda(r^L, r)(1-\gamma)D, \rho] \qquad (17^J)$$

Note that this equilibrium condition hides many salient aspects of the credit market in ρ and only explicitly captures the bank-lending channel under the *ceteris paribus* clause of no alteration in the riskiness and quantity of credit.

Determination of aggregate demand

For comparative static analyses, the IS equation (7), the LM equation (13) and the loan market CC equation (17) specify the three equations needed for deriving aggregate demand.[36] They can be solved for ρ, r (or for R, given π^e) and aggregate demand y^d in terms of the values of the exogenous monetary policy variables γ and M^s, and the fiscal policy variables.

[35] Bernanke and Blinder (B&B) (1988) do not derive the demand for loans but instead assume a priori the loan demand function to be:

$$L^d = L^d(\underset{-}{R}, \underset{+}{R^L}, \underset{+}{y})$$

Hence, the B&B loan market equilibrium condition is:

$$L^d(r, r^L, y) = \lambda(R^L, R)(1 - \gamma)D$$

36 This model of aggregate demand would be very similar to that of Bernanke and Blinder (B&B) (1988) if credit were defined as bank loans and our emphasis on quantity rationing and the layered structure of credit and financial institutions, designated by ρ, were ignored.

For this purpose, first combine the IS and credit market equilibrium equations by eliminating R^L to derive the "IS–CC equation." The general form of this equation is:

$$y^d = y^d(r, P, (1-\gamma)D; g, \rho) \tag{18}$$

where g stands for the fiscal policy variables. In (18), $\partial r/\partial y < 0$ since a rise in r reduces investment, as for the IS curve, so that the IS–CC curve has a negative slope in the (r, y) space. However, note that while the IS curve is independent of the money market, the IS–CC curve shifts with changes in the monetary base M0, the currency ratio *C/D* and the required reserve ratio γ, which together determine deposits D. Shifts in this curve can also occur because of shifts in M0 and ρ.

If credit and bonds are perfect substitutes[37] for either the borrowers or the lenders, so that $R^L = R$ and the supply of credit is not rationed on a basis other than their rate of return, or if commodity demand does not depend on the loan interest rate, the IS–CC equation degenerates to the IS one, so that its combination with the LM equation makes the model the standard IS–LM curve for determining aggregate demand. However, intuitive and empirical information on the economy seems to indicate that these conditions are not likely to be met for the financially developed economies,[38] and even less so for the financially underdeveloped economies. Further, our model assigns distinctive short-run roles to credit and bonds in the production sector, so that they cannot be perfect substitutes in our model.

Determination of output

In the Blinder and Bernanke (B&B) (1988) analysis, variations in aggregate demand cannot change output unless there are imperfect price adjustments, such as when prices are sticky, irrespective of whether bonds and credit are perfect substitutes or not. But if the economy has imperfect price adjustment and if bonds and credit are not perfect substitutes, the B&B model implies different effects of monetary policy on aggregate demand, and consequently on output, from those in the IS–LM model. This also applies to our model.

Given our assumption for comparative static analysis that $\pi^e = 0$, $R = r$ and $R^L = r^L$, the four endogenous variables in these three equations are y, P, r and r^L. Combining the IS, LM and CC equations to eliminate r and r^L, we end up with the aggregate demand equation:

$$y^d = y^d(\underset{-}{P}; m, g, \rho) \tag{19}$$

where m is a monetary policy variable (money supply and/or interest rate) and g is a fiscal policy one. The commodity supply side of our model has a production sector in which credit and bonds play different roles: changes in credit change the working capital of firms. Its earlier analysis gave the commodity supply function as:

$$y^s = y^s(\underset{-}{r^L}, \rho) \tag{20}$$

[37] The assessment of the literature on this point by Kashyap and Stein (1994) is that loans and bonds are not perfect substitutes.

[38] For this assessment of the empirical information, see Kashyap and Stein (1994).

Therefore, equilibrium in the commodity markets yields the output equation as:

$$y = y^{\mathrm{d}}(P; m, g, \rho) = y^{\mathrm{s}}(r^{\mathrm{L}}, \rho) \tag{21}$$

The credit market equilibrium condition derived earlier was:

$$P \cdot L^{\mathrm{d}}(r^{\mathrm{L}}, y) = L^{\mathrm{s}}[\lambda(r^{\mathrm{L}}, r)(1-\gamma)D, \rho] \tag{22}$$

The output equation (21), the CC condition (22) and the LM condition (13) can be combined by eliminating r and r^{L}. This yields the output produced as:

$$y = y((1-\gamma)D; g, \rho) \quad \partial y/\partial((1-\gamma)D) > 0 \tag{23}$$

Note that in our model the monetary policy variables determine the supply of loans, which enters as one of the components of working capital in the indirect production function. Hence, there is a positive relationship between y and $(1\ \gamma\ )D$, so that an expansionary (contractionary) monetary policy increases (decreases) output.[39] Output also depends on ρ, our proxy for the riskiness and quantity of credit. If we interpret an increase in ρ as a loosening of credit rationing, y would depend positively on ρ.

Impact of monetary and fiscal policies

Impact of monetary policies on credit conditions

First, consider the impact of a tight monetary policy that increases both the bond and credit interest rates, thereby decreasing investment and the aggregate demand for commodities. Such an effect on aggregate demand also occurs in the standard IS–LM analysis. In addition, in our model, the fall in aggregate demand worsens business conditions for firms, thereby reducing their net worth and increasing the riskiness of credit, so that banks reduce the proportion of the portfolio that they wish to allocate to credit. This action will reduce the credit supply and further increase the credit interest rate. However, a sufficient shift of the banks' portfolio to bonds could end up increasing the banks' demand for bonds, so that the overall effect on the bond rate would become negative.[40] A smaller shift would still leave the bond rate higher, but moderate its increase. Therefore, in our money–bonds–credit (M–B–C) model, a tight monetary policy could lead to a smaller increase in the bond rate or a fall in it, while raising the credit interest rate. In any case, the spread between the credit interest rate and the bond rate will increase.

The impact of a contractionary monetary policy on short-run output proceeds through two channels. One is through the restriction in aggregate demand, which decreases output and prices if there are market imperfections of the type emphasized by the new Keynesians

[39] The limit to the expansionary impact occurs if monetary policy increases the supply of working capital beyond its maximum demand as output increases.

[40] Bernanke and Blinder (1988) cite March to July 1980 as a period during which a tight monetary policy in the USA reduced the government bond rate. Bernanke (1983) attributed the length of the Great Depression to a reduction in the supply of loans arising from the increase in their riskiness and banks' demand for increased liquidity/reserves due to an increase in the possibility of runs on them.

(see Chapter 15), but only decreases the price level if there are no market imperfections. The other channel of impact proceeds through the policy's impact on working capital: the contractionary monetary policy decreases the availability of credit and reduces the working capital of firms, which reduces their output. Therefore, the contractionary monetary policy reduces short-run output whether there are market imperfections or not, but the reduction in output is likely to be greater if there are market imperfections.

An expansionary monetary policy will improve business conditions by increasing aggregate demand and the sales made by firms. This will make credit to firms less risky and tilt the banks' portfolio composition towards credit from bonds. The expansionary monetary policy will lower both the credit and bond rates, while the portfolio composition shift will lower the credit interest rate further and moderate the fall in the bond rate. These will increase aggregate demand in the economy and would decrease output under a new Keynesian context of market imperfections. In addition, the greater availability of credit will increase the working capital of firms and, unless they were already at the maximum level needed for the current level of production, would increase output.

Impact of fiscal policies

An expansionary fiscal policy, with deficits financed by bond issues, will increase aggregate demand in the usual manner in the IS–LM and AD–AS analysis. This will increase output if there are market imperfections. The expansionary fiscal policy will also raise the bond rate, thereby causing a readjustment of banks' portfolio such as to decrease their credit supply. This has two effects. One is to raise the credit interest rate; the portfolio shift to credit from bonds will further increase the bond rate while moderating the increase in the credit interest rate. The second effect of the decrease in the supply of credit, and of the increase in the credit interest rate, is to reduce the working capital of firms, which will reduce output. Hence the expansionary fiscal policy has two contrary effects: in the presence of market imperfections, the increase in output due to the increase in aggregate demand, and the fall in output because of the decrease in working capital. The net effect is indeterminate when there are market imperfections. However, in the absence of such imperfections, the net effect will be negative: the expansionary fiscal policy will decrease output. The impact of a contractionary fiscal policy will be the opposite.

Impact of other causes of variations in credit supply or demand

Suppose that the riskiness of credit increases because of a negative shock, as would occur in a downturn in the economy, and adversely affects firms' ability to repay credit provided to them. This would reduce banks' supply of credit, so that, given bank deposits, the banks' demand for bonds would rise.[41] While the decreased supply of credit will raise the credit interest rate, the increased demand for bonds will lower the bond rate. The latter, as in the IS–LM model, will increase aggregate demand, which could, if there are new Keynesian market imperfections, increase output. However, the decrease in credit will reduce firms' working capital and output. Hence the net effect on output is indeterminate in the presence of market imperfections, but would be negative otherwise.

[41] The banks' demand for free reserves may also rise, though, in many modern economies, changes in free reserves tend to be of minor significance.

A negative productivity shock would reduce productivity and demand for both labor and working capital. It also affects firms' ability to repay their loans and their borrowing through short-term corporate paper, so that it increases their risk. Therefore, a productivity shock would decrease output more in the short-run than in the long-run because of its impact on the short-run supply of credit.

Relative size of impact on small versus large firms

Our model relates credit supply and the credit interest rate to the working (and possibly physical) capital of firms and thereby to their short-run production/output of commodities. On this basis, assume that a decrease in the supply of credit reduces the working capital of the borrowing "small" firms and their production more than it does that of relatively large firms which have prearranged their short-run working capital needs through issues of bonds and retained earnings at the beginning of the current short-run. Further, there would be a flight to quality (less risky loans) as credit supply decreases, so that there would be differential effects; the effect will be more severe on the more credit-dependent small and intermediate-sized firms.

If the number of firms that rely on credit to facilitate production is significant enough in the economy, a decrease in the supply of credit will produce a fall in the output of the borrowing firms and, therefore, of the economy. A corresponding decrease in the demand for bonds will not elicit a similar impact in the short-term since this would not reduce funds already raised through bond issues and so will not reduce the working capital of firms. It may, however, affect the ability and cost of raising funds for the future through subsequent bond issues. Therefore, in the short-run, the impact on production and employment would be greater and occur faster for a decrease in the credit interest rate and/or the supply of credit than of a corresponding fall in the bond rate or decrease in the demand for bonds.

Instability in the money and credit markets and monetary policy

Controllability of credit supply by the central bank

Assuming that the credit market is an important independent source of effects on real output, the relevant question from the perspective of central bank policy is whether or not it can control the demand or supply of credit in some manner or control the credit interest rate. Since credit is fungible between bank loans, trade credit and short-term commercial paper, and trade credit has a certain amount of elasticity, a mild restrictive monetary policy that reduces bank loans somewhat may just be offset by an extension of trade credit, without any impact on working capital. This would eliminate the direct impact of the contractionary policy on output supply. Consequently, only the demand channel of monetary policy effects would be relevant. By comparison, a contractionary monetary policy severe enough to reduce working capital overall would have effects on both the supply side and the demand side. This implies a non-linear impact of monetary policy on output. The converse should roughly apply for the effects of expansionary monetary policies.

Central banks in several countries, such as the USA and Canada, pursue monetary policy through changes in the monetary base or/and the bond interest rate but do not pursue policies that directly affect the demand or supply of credit or the credit interest rate. Indirect central bank control of the supply of credit by banks through changes in the monetary base, brought about by open market operations, occurs through the latter's impact on bank deposits.

To the extent that the impact of a decrease in the monetary base on bank deposits is neutralized by banks through borrowing in the bond markets, such as by the sale of the banks' own short-term bond issues (such as certificates of deposit in the USA) to the public, the central bank's control over the loan supply by banks through changes in the monetary base will be diluted.

General monetary policies are likely to differentially affect both the credit and the bond markets in a manner determined by the economy and the lenders' portfolio decisions. They cannot be selectively targeted to the credit market alone. Therefore, their pursuit would be a somewhat undiscriminating instrument for manipulating the credit supply and/or its interest rate alone. However, central banks in many countries do pursue distinct policies affecting the credit market. Among such policies are selective ones that set the interest rates on credit or specify the allocation of banks' portfolio, or order specified increases in credit to specific sectors such as agriculture, exports and housing. To illustrate, on the purchase of durable consumer goods financed through installment credit,[42] some central banks set the down/initial payment and the period over which the loan has to be repaid.

Short-run versus long-run effects in Keynesian and neoclassical models

The preceding analysis has been short-run. In this short-run, changes in credit change firms' working capital, which alters output and employment. They also change aggregate demand through their impact on the credit and bond rates and, through them, on investment and consumption. If the economy is Keynesian, with imperfectly competitive firms setting prices and with some form of price rigidity (e.g. due to menu costs under a Calvo-type price adjustment mechanism), a change is demand produces a change in both output and prices (Clarida *et al.*, 1999). Therefore, the Keynesian economy has two mechanisms, which together produce a change in output following a change in supply of loans. But if the economy is neoclassical, with perfect competition in all markets, changes in aggregate demand bring about a change only in the price level, so that a change in credit supply would alter output only through its impact on working capital. Therefore, in the short-run, a restrictive monetary policy would produce a decline in output in both the neoclassical and Keynesian models, but with a greater decline in the Keynesian case.

In the long run, changes in credit are only one mode of adjusting working capital, since retained earnings and bonds can also be used for this purpose. The economy will then have the amount of working capital required for output at the full-employment level, and changes in credit will not affect output. Further, the short-run price and nominal wage rigidity adduced by Keynesians will not exist in the long-run, so that changes in aggregate demand will not cause output to differ from its full-employment level. Hence, there is long-run neutrality of loans, just as of the money supply, whether the model is neoclassical or Keynesian and irrespective of whether a distinction is made between credit and bonds.

The financial instability hypothesis

Hyman Minsky (1986), along with other economists in the post-Keynesian tradition, argues that the financial sector is inherently unstable and possesses the capacity for destabilizing the real economy. The following provides some flavor of these arguments. As a boom progresses, firms increase their investment spending. To finance it, they increase their debt relative to their

[42] Such policies go under the name of hire-purchase or installment-credit controls.

income flows, with their debt instruments becoming increasingly speculative. The financial markets are willing to absorb these during the boom as firms' profits rise and the financial markets, in an atmosphere of euphoria/exuberance and of "easy money," assess the risk of bond holdings to be relatively low. This euphoria also reduces the degree of risk aversion of lenders. The resulting disregard for risk leads to an expansion of credit, some of it extremely risky, which makes the financial sector vulnerable to a variety of shocks,[43] some of which can trigger a financial crisis. During this crisis, there occurs a reassessment of the riskiness of bonds and credit, as well as increasing risk aversion, so that the supply of funds to the credit market falls and their cost rises. The consequence is tighter credit, with access to funds becoming closed to some firms and reduced for others. This forces firms that rely on such funds for working capital and for financing short-term investment, e.g. in inventories, to cut back on them. The crisis may also be accompanied by cutbacks in consumer expenditure. These bring about reductions in production and investment in the economy as a whole. The resulting fall in profits intensifies the trend toward tighter credit and rising interest costs. The boom-time euphoria gets replaced by pessimism and panic during the downward spiral.

This story can be embellished by that on the existence of herds (Chari and Kehoe, 2002) in financial markets. In the absence of hard information on the future profitability or solvency of borrowers, in some cases, though not in others, an increasing perception of risk by some lenders is followed by a similar reassessment of risk by other lenders, so that the funds made available to borrowers as a whole decrease and trigger a decrease in production. Conversely, in some cases, though not in others, a decreasing perception of risk and aggressive lending by some financial institutions evokes a similar response from others, causing a general stampede toward easier credit, resulting in production increases. Business and consumer confidence on the future course of the economy, aggregate demand and job availability, are also quite susceptible to the herd phenomenon.

Credit channel when the bond interest rate is the exogenous monetary policy instrument

Chapter 13 argued that the central bank might use the interest rate as its primary instrument of monetary policy and might follow a Taylor rule for this purpose. This changes the preceding analysis by making the supply of money endogenous. In this context, suppose that the central bank pursues an expansionary monetary policy by lowering the bond interest rate. This has two immediate effects. One, to ensure equilibrium in the financial markets, the central bank has to ensure an adequate money supply by increasing the monetary base. Two, the decline in the bond rate makes credit more attractive in banks' portfolios, which causes banks to increase their supply of credit for a given money supply. This substitution of credit for bonds lowers the credit interest rate, which induces firms to take more loans. Hence, the expansionary monetary policy increases credit in the economy, which increases working capital. Therefore, the effects of an expansionary monetary policy implemented through a decrease in the bond rate are similar to those analyzed above for an expansion of the money supply.

[43] Examples of such shocks include interest rate increases by the central bank in order to fight inflationary pressures, unexpected default by some borrowers, lower realized profits by firms than were expected, etc.

However, note that the impact on working capital is more directly through the money supply and less through interest rates.

The informal financial sector and financial underdevelopment

All economies also have an informal financial sector. Borrowers in this market usually do not have access to credit from the organized financial institutions such as banks, bond and stock markets, etc., but have to borrow from "moneylenders." This informal financial sector shares many of the characteristics of the "organized loan market" mainly operated by banks: normally, a particular borrower has access to credit from one or a few lenders, the credit is based on personal knowledge of the borrower and can often be recalled on demand or short notice. Therefore, for our analysis, the loan market can be taken to encompass credit given in both the organized and the informal financial sectors. Adding in the latter makes the credit market much more significant in economies with a substantial informal financial sector.

The existence of credit markets with quantity rationing means that the amount of credit extended is likely to be more closely related to the money supply than the credit interest rate is to the bond rate. This is especially so in financially underdeveloped economies in which the interest rates charged in the informal financial sector are likely to be only loosely related to the bond rates in the organized financial markets. In this case, the best monetary policy instrument is more likely to be the money supply rather than the interest rate. Consequently, even if the interest rate proves to be the more appropriate instrument in developed economies, as some estimations of the Taylor rule show (see Chapter 13), it need not be so for the developing economies.

Bank runs and credit crises

The financial system is prone to several special characteristics. Among these is contagion and runs, which create "sympathetic co-movements" in asset prices and liquidity.

Contagion is the spread of similar sentiments across financial institutions and instruments. For example, at the level of institutions, a widespread belief that a bank is about to fail leads to suspicions of the vulnerability of some other banks. At the level of assets, for example, a decline in the share prices of a firm in a particular industry leads to declines in the share prices of other firms. Contagion, therefore, leads to sympathetic co-movements across the institutions and assets of the financial system, and can be empirically observed in the self-reinforcing character of declines in asset prices. They also lead to a "herd movement" in which traders and individual investors rush to avoid losses by selling their asset holdings.

A run occurs because financial institutions typically hold assets with a longer maturity than their liabilities. In the case of banks, banks' liabilities are mainly withdrawal of deposits on demand, while their assets mainly consist of short-term bonds, loans, mortgages, etc. If there is sudden widespread withdrawal of deposits (a run), the bank's assets cannot be liquidated soon enough to meet the withdrawals, thereby forcing losses if the assets were liquidated at short notice, or insolvency for the bank. In addition, contagion can spread a run on one bank to other banks. Similar problems can afflict other financial institutions.

Financial systems have evolved several mechanisms to counter runs and contagion. Among these are the markets for trading reserves on an overnight basis (the federal funds market in the USA), the lender-of-last-resort function of central banks (see Chapter 11), insurance of bank deposits, and banking supervision to ensure proper financial practices. However, while these reduce the possibility of runs, runs can and still do occur even in financially developed economies.

The relevance for the credit channel of runs on financial institutions and contagion among them is that their occurrence affects the availability and cost of credit in the economy, which can have real effects. Consequently, one of the roles of the central bank is to prevent their occurrence and, if they do hit, mitigate their effects on the financial system and the economy.

Empirical findings

As discussed in Chapter 2, monetary policy can affect aggregate demand through a number of transmission channels or mechanisms. Of these, the direct channel operates through the spending of excess money balances directly on commodities, while the indirect channel operates through a change in the interest rate, which alters investment. The indirect transmission channel is embodied in the IS–LM framework. The impact of money on output through these two channels has been labeled by some authors the "money view," as against the credit channel. The open economy also allows monetary policy to affect the balance of payments, changes which impact on domestic expenditure and output.

Among studies that did not find a significant independent impact of the loan channel on output are those of Oliner and Rudebusch (1996) and Driscoll (2004). Driscoll uses panel data from states within the USA to examine the impact of bank lending on output. He finds that shifts in money demand have large and statistically significant effects on the supply of bank loans, so that monetary policy can affect the latter. However, bank loans only have small, statistically insignificant, impact on output. Therefore, the bank lending channel is not a significant, independent contributor to the impact of monetary policy on output. Note that most such findings relate to the impact of bank loans, not of credit as a whole, on output. This chapter implies that changes in the overall credit supply affect output. If shifts in bank loans were offset by responsive shifts in other forms of credit, which are trade credit and short-term commercial paper, shifts in bank loans would not affect output.

Further, the regression of output on the money supply yields an estimate of the total impact of shifts in it on output. Since money supply and credit often tend to move together, in a regression of output on money the regression coefficient of money would tend to include within it at least some of the impact of changes in credit on output. If the estimation is then extended to include credit, in addition to money, as an explanatory variable, the marginal increase in predictive power may not prove to be significant. This is especially so for bank credit since bank credit and bank deposits move together because they reflect different sides of the banks' balance sheet. This makes it difficult to find a separate impact of bank loans in addition to that of money (King, 1986; Romer and Romer, 1990; Ramey, 1993; Walsh, 2003, Ch. 7). However, as the asset-backed security crisis in the USA in 2007 illustrates, shifts in the riskiness of credit and its perception and in the supply of credit can have distinct effects on output. Significant shifts of this kind tend to occur rarely, so that their impact may be more visible in detailed case studies of episodes in which the movement of credit diverged from that of money supply.

On this issue, Bernanke and Blinder (1988) compared the correlations between the growth rates of (nominal and real) GNP and money with those between GNP and credit[44] for the USA.

[44] The money measure used was M1. Credit was "the sum of intermediated borrowing by households and businesses" from Flow of Funds data.

The correlations of GNP with money were higher for 1953:1–1973:4 than for credit but lower for 1979:4–1985:4. They also used simple least squares to estimate money and credit demand functions incorporating a partial adjustment model. While they could not check for parameter instability, the variance of the residual of the money demand equation was smaller for 1974:1–1979:3, but larger for 1979:4–1985:4, than that of the credit demand equation. They concluded that money demand shocks became relatively greater in the 1980s. While such evidence is suggestive rather than compelling and does not establish the direction of causality, Bernanke and Blinder claim that it may now be better for central banks to target credit rather than money.

Among studies supporting a distinctive lending channel, Kashyap and Stein (1993, 2000)[45] provide evidence supporting the distinction between loans and bonds and show that neither banks nor firms were indifferent between them. Consequently, changes in the loan interest rate or the loan supply have a different impact on aggregate demand than a corresponding change in the bond rate and the demand for bonds. In addition, this chapter's analysis views credit as providing part of the working capital needs of small and intermediate firms. A decrease in the availability of credit relative to their demand forces a decrease in their working capital and in their production, so that credit has a distinctive impact on both aggregate demand and supply in the economy. Bernanke (1986) finds that lending shocks have a sizeable effect on aggregate demand. Bernanke and Blinder (1992) report that as banks adjust their credit in response to monetary shocks, the decline in credit reduces output growth. Gertler and Gilchrist (1994), among many other studies, report that monetary tightening, such as by an increase in the interest rate, by the Fed often leads to a decline in bank loans to small firms and that the inventory investment of small firms is especially sensitive to such changes.

However, even if credit is an endogenous variable and changes in it are not sufficiently exogenous to changes in money or the bond rate, the credit structure of the economy shapes the dynamic response of real economic activity to monetary policy, so that it could still play an important role in the way that the impact of financial disturbances is propagated in the economy. Intuition indicates that the composition of spending among industries, firms and consumers does depend on the credit and bond structure of the economy. The compositional relevance of the credit channel can be established by examining shifts in the composition of economic activity among industries and firms in response to monetary policy shocks, and is supported by several empirical studies (see Kashyap *et al.*, 1993; Gertler and Gilchrist, 1994; Lang and Nakamura, 1995; Ludvigson, 1998). To illustrate, Ludvigson reports that a tight monetary policy reduces bank consumer loans, which reduces real consumption expenditures, so that the composition of aggregate expenditures changes.[46]

Since different economies, especially financially developed as opposed to undeveloped economies, can have different bond and credit structures, international comparisons of the overall impact of monetary policy on output may provide some, though indirect, evidence on the relative importance of the credit channel. More direct evidence would have to come from the compositional effects of monetary policy among industries, firms and consumers.

[45] Kashyap and Stein (2000) demonstrate that lending by small banks is relatively more sensitive to monetary policy. However, Ashcraft (2006) and Ashcraft and Campello (2007) report that the *average* impact of monetary policy on lending by banks is not statistically significant, while changes in the creditworthiness of borrowers, and therefore the balance sheet channel, do affect the response of bank lending to monetary policy.

[46] Note that the empirical evidence of the compositional effects does not provide much guidance on the magnitude of the *quantitative* importance of credit effects in the overall impact of monetary policy on output.

These effects should differ among countries and be most visible in comparisons between the financially developed and the undeveloped economies.

Conclusions

The extensive consideration of market imperfections in macroeconomic modeling in recent years has extended to their role in financial markets. The imperfections in credit markets arise from adverse selection, moral hazard and monitoring and agency costs, and financial underdevelopment. Consequently, bonds and credit, which includes loans, are not perfect substitutes, so that their different characteristics can be better accommodated byclassifying financial assets into three categories, money, bonds and credit, rather than two, money and bonds, which is the pattern of the IS–LM and IS–IRT models. The addition of information imperfections and credit help to explain many of the distinctive features of financial markets and their impact on the real economy.

This chapter relies on the addition of two innovations to the IS–LM and IS–IRT models of aggregate demand. One of these is to use information imperfections to draw a distinction between bonds (including equities) and credit. The other is the use of an indirect production function, which has working capital as an input since it facilitates the purchases of inputs (labor, raw materials and intermediate goods) bought prior to production and sales. Working capital usually comes from retained earnings, bonds (including stocks) and credit (including loans). Of these, this chapter's model simplified the analysis by assuming that the amount of working capital obtained by the firm from retained earnings and bonds does not vary in the short-run, but can be varied by the firm in the long-run.

Expansionary monetary and fiscal policies have somewhat different effects on output in our model. An expansionary monetary policy increases credit supply and reduces the credit interest rate, as well as increasing aggregate demand. An expansionary fiscal policy increases output in a new Keynesian model with market imperfections, but not in a neoclassical one (without market imperfections). The monetary policy, in our model, but not in the new Keynesian or neoclassical ones, increases the amount of working capital and output. Therefore, the expansionary (contractionary) monetary policy causes an unambiguous increase (decrease) in output.

Imperfections in financial imperfections imply that the decrease in credit in financial panics or runs will add to the impact of any decrease in the money supply on output. This reasoning implies that the contractionary impact of the financial sector in the Great Depression of the 1930s occurred not only through the decrease in the money supply but also through much more intense credit rationing, which worsened the fall in output and extended its duration.

Note that our definition of loan suppliers included moneylenders among banks. Developing economies have a larger proportion of firms that rely on loans rather than bond issues for their working capital needs. They also have a relatively large informal financial sector. Our analysis shows that the impact on production of a fall in the supply of loans is likely to be more intense and speedier in developing economies than in financially developed ones.

Since a change in the money supply is positively related to changes in the amount and cost of credit, the credit channel intensifies the impact of monetary policy on aggregate demand beyond that encompassed in the "money view," i.e. monetary transmission effects on aggregate demand through the bond interest rate. Further, the overall effect of money on output now includes, in addition to effects through aggregate demand, a working capital effect, since credit, but not money itself, is a component in the proposed short-run production function.

There are vertical layers among the institutions that provide credit. The link between money supply and these credit layers tends to be imprecise, so that the supply of credit cannot be controlled accurately enough by monetary policy. Hence the central bank may only be able to moderate, but not fully offset, the effects of a credit crisis, arising say from a heightened perception of the riskiness of lending to borrowers in credit markets, on output and employment, through an expansionary monetary policy. The credit channel, therefore, can dilute the central bank's control of the economy, especially during a credit crunch or credit euphoria.

Appendix A

Demand for working capital for a given production level in a simple stylized model

To illustrate the firm's demand for working capital and the differentiation between it and the firm's money holdings, we adapt Baumol's (1952) inventory analysis of the demand for money as a medium of payments in the presence of an alternative financial asset. The two financial assets in the following analysis are money and loans/credit. We assume that the firm purchases its input (labor, raw materials and intermediate inputs) and has to pay for them in an even stream during the period. These payments are made in advance of production and sales, with sales assumed to occur only at the end of the period. For simplification, the values of the purchases and sales revenues are assumed to be identical. The firm finances its purchases through a loan, arranged from a bank at the beginning of the period in the form of an overdraft. The loan to the firm is repaid at the end of the period from its sales revenue.

Let the total cost of inputs (and the sales revenue) equal $\$Y$. The firm is assumed to withdraw from its overdraft the amounts needed for its payments in an even stream through the period, with the amount drawn each time being $\$z$ and the number q of withdrawals equal to Y/z. Since the firm will withdraw z at the beginning of the period, after an interval equal to $1/q$ of the period, after another interval equal to $1/q$, and so on, the average amount withdrawn and spent on purchases is:

$$K^{w,d} = z + z(q - 1)/q + z(q - 2)/q + \cdot \cdot \cdot + z/q$$

$$= (q + 1)z/2 = ½Y + ½z \qquad (24)$$

Since $qz = Y$, the firm's average money balances will equal $M/2$, so that:

$$K^{w,d} = ½Y + M \qquad (25)$$

Hence, the demand for working capital is greater than for money balances. In real terms, this demand becomes:

$$K^{w,d}/P = ½y + m \qquad (26)$$

where $y = Y/P$ and $m = M/P$.

Equation (25) specifies the *maximum* amount of working capital needed by the firm to finance a given amount of purchases of inputs. The firm can economize on its working capital by reducing its output or/and by diverting some of its labor to save on working capital. For a given maximum need for working capital, the firm may also be able to substitute to some extent trade credit for loans, but this will also usually involve some diversion of labor to make alternative trade credit arrangements with the suppliers and buyers from the firm. This possibility is investigated in Appendix B.

In the long run, the firm can provide for its working capital needs through bonds, retained earnings, trade credit and loans. Of these, in the short-run, credit is the only component that is taken to be variable in this chapter.

Appendix B

Indirect production function including working capital

The firm's demand for working capital is the average amount of the "funds" that it wants to hold in order to carry out its purchases of labor services and other inputs (raw materials and intermediate goods). The following analysis justifies the appearance of working capital as an input in the production function and the determination of its optimal amount through profit maximization.

Assume that the firm's output depends on its physical capital and the part of its employment that it uses directly as an input in production. However, it has to divert some of its workers to carrying out transactions involving purchases of inputs and the sale of its output. If the firm held only a small and relatively inadequate amount of working capital, it would have to employ workers in juggling its working capital to carry out the required transactions of purchase and sale of commodities, as well as payments to workers. Firms usually use a mix of trade credit and their own working capital.[47] However, arranging and using the former implies a relatively higher amount of labor time than if the firm holds adequate working capital. The use of working capital, therefore, allows the firm to economize on the workers it has to divert to making payments.

The preceding arguments yield the representative firm's production function as:

$$y = y(n_1) \quad \partial y/\partial n_1 > 0, \partial^2 y/\partial {n_1}^2 < 0 \tag{27}$$

where y is the firm's output and n_1 is the amount of labor directly involved in production. Total employment by the firm is n_1 n_2, where n_2 is the labor used in the payments and receipts processes so that: +

$$n_1 = n - n_2 \tag{28}$$

where n is total employment. For given n, $\partial n_1/\partial n_2 < 0$.

For the labor used in carrying out exchanges, and using the firm's output y as the proxy for the number of transactions involved in purchasing inputs, the use of n_2 workers is taken to be given by:

$$n_2 = n_2(k^w, y) \tag{29}$$

[47] Guariglia and Mateut (2006) report on the existence of substitutability between trade credit and credit (loans)

at the micro level for the UK, and that this substitution weakens the loan channel.

where $k^w (\equiv K^w/P)$ is the firm's real working capital, $\partial n_2/\partial k^w \leq 0$ and $\partial n_2/\partial y > 0$. The specific form of $n_2(.)$ would depend on the trading and payments technology of the economy and would shift with that technology. Innovations in the financial system, such as the use of direct deposit of salaries into the workers' accounts and payments to suppliers by electronic transfers, would reduce the demand for real balances for transactions associated with a given level of output and shift the transactions technology function. It will also depend on the availability and the flexibility of trade credit.

From (27) to (29),

$$\frac{\partial y}{\partial k^w} = \frac{\partial y}{\partial n_1}\frac{\partial n_1}{\partial n_2}\frac{\partial n_2}{\partial k^w} \geq 0$$

A specific form of (29) is the proportional expression:

$$n_2/y = \varphi(k^w/y) \tag{30}$$

where $\varphi' \equiv \partial\varphi/\partial(k^w/y) \leq 0$. For this function, the firm reaches "saturation" in real balances relative to its output when $\varphi' = 0$, which will occur when the firm holds the maximum amount of working capital, as derived in Appendix A. From (27) to (30),

$$y = y(n - y \cdot \varphi(k^w/y)) \tag{31}$$

which can be rewritten as the indirect production function:

$$y = f(n, k^w) \tag{32}$$

where $\partial y/\partial k^w$ 0. Hence, the use of working capital by the firm increases its output, with its marginal product being positive up to the saturation point $k^{w,\max}$. Up to this point, the use of working capital allows the firm to reduce the labor allocated to payments, thereby increasing the labor allocated directly to production, which increases the firm's output for a given amount of employment.

Profit maximization and the optimal demand for working capital

The firm is assumed to operate in perfect competition in all (output and input) markets and to maximize profits. Its profits are given by:

$$M = PF(n, k^w) - Wn - R^L \cdot Pk^w - F_0 \tag{33}$$

where M is profits, P is the price level and F_0 is fixed cost, which includes the firm's commitment to pay interest on its existing bonds. n is employment, W is the nominal wage rate and k^w is the amount used of real working capital.

The first-order conditions for maximizing profits with respect to n and k^w, are:

$$P \cdot \partial F/\partial n - W = 0 \tag{34}$$

$$P \cdot \partial F / \partial k^{\mathrm{w}} - P \cdot R^{\mathrm{L}} = 0 \tag{35}$$

Solving (34) and (35) yields the demand functions:

$$n^{\mathrm{d}} = n^{\mathrm{d}}(w, R^{\mathrm{L}}) \tag{36}$$

$$k^{\mathrm{w,d}} = k^{\mathrm{w,d}}(w, R^{\mathrm{L}}) \tag{37}$$

where *w W/P*. The supply of output is given by substituting (36) and (37) in the indirect production function (32). Its functional form is:

$$y = f(w, R^{\mathrm{L}}) \tag{38}$$

With $\pi^{\mathrm{e}} = 0$, this function becomes $y = f(w, R^{\mathrm{L}})$.

Demand for credit

The short-run nominal demand for credit is given by:

$$L^{\mathrm{d}} = Pk^{\mathrm{w,d}}(w, R^{\mathrm{L}}) - K^{\mathrm{w\#}} \tag{39}$$

where $K^{\mathrm{w\#}}$ is the nominal amount of working capital prearranged through bonds and retained earnings. Short-run variations in $Pk^{\mathrm{w,d}}$, therefore, produce corresponding variations in the short-run demand for loans. Hence:

$$L^{\mathrm{d}}/P = \psi(w, R^{\mathrm{L}}; K^{\mathrm{w\#}}) \tag{40}$$

Summary of critical conclusions

- ❖ Adverse selection, moral hazard and monitoring and agency costs imply quantity credit rationing, in addition to rationing by the interest rate charged.
- ❖ Imperfections in financial markets lead to a distinction between bonds and credit.
- ❖ Firms need working capital to facilitate the purchases of inputs and sale of output, so that it becomes an argument of their indirect production function.
- ❖ The assumption that the funds raised through bonds (including stocks) are given in the short-run, while those raised through credit (especially trade credit and bank loans) are variable, allows monetary policy to change the working capital of firms and its cost, so that it has a direct impact on the supply of commodities in the economy.
- ❖ The credit channel is one of several channels that determine the magnitude and lags in the impact of monetary policy on output.
- ❖ It is difficult to empirically estimate the marginal impact of the lending channel on the economy’s relationship between money and output. However, its impact can be established convincingly in terms of its effects on the composition of expenditures among small and large firms, and between firms and households, and of output among industries.
- ❖ The credit channel is likely to be relatively more important in financially

Review and discussion questions

1. Discuss the main characteristics of money, bonds, credit and equities/stocks in actual financial markets. What is gained and what is lost by having a macroeconomic model

with only two financial assets, money and bonds, with the latter including credit and equities?

2. Discuss the sources of imperfections in credit markets and their role in quantity and interest rate rationing.
3. Specify a model of aggregate demand with three financial assets, money, bonds and credit.
4. Given your model of aggregate demand when there are three financial assets, money, bonds and credit, and the standard classical production function, would disturbances in the loan market cause changes in output and employment? Your answer should present the determination of output and employment in this model.
5. Given your model of aggregate demand when there are three financial assets, money, bonds and credit, show how the new Keynesian model allows shifts in the credit market to alter output and employment.
6. What is working capital? Justify its inclusion in a production function and show the impact of a decrease in working capital on the supply of short-run output and employment.
7. Given your model of aggregate demand with three financial assets (money, bonds and credit) and the indirect production function with working capital (but without the new Keynesian Phillips curve), how can monetary policy produce an increase in short-run output and employment? Can it do so in long-run output and employment? Discuss.
8. Given your model of aggregate demand with three financial assets (money, bonds and credit) and the indirect production function with working capital and the new Keynesian Phillips curve, what are the effects of a contractionary monetary policy on output and employment in the short-run? What are its effects in the long run? Discuss.
9. Discuss why the credit channel is likely to be more important in financially developing economies than in developed ones, and discuss its implications for the choice between the money supply and the interest rate as the appropriate monetary policy instrument.
10. Discuss the significance of the credit channel in changing aggregate demand and output. What limitations, if any, on this significance are imposed by the addition of the expectations-augmented Phillips equation? What limitations, if any, on this significance are imposed by the addition of short-run money neutrality?

References

Akerlof, G. “The market for ‘lemons’: quality uncertainty and the market mechanism.” *Quarterly Journal of Economics*, 84, 1970, pp. 488–500.

Ashcraft, A.B. “New evidence on the lending channel.” *Journal of Money, Credit and Banking*, 38, 2006, pp. 751–75.

Ashcraft, A.B. and Campello, M. “Firm balance sheets and monetary policy transmission.” *Journal of Monetary Economics*, 54, 2007, pp. 1515–28.

Baumol, W.J. “The transactions demand for cash: an inventory theoretic approach.” *Quarterly Journal of Economics*, 66, 1952, pp. 545–56.

Bernanke, B.S. “Nonmonetary effects of the finanical crisis in the propagtion of the Great Depression.”

American Economic Review, 73, 1983, pp. 257–76.

Bernanke, B.S. “Alternative explanations of the money–income correlation.” *Carnegie-Rochester Conference Series on Public Policy*, 25, 1986, pp. 49–99.

Bernanke, B.S. "Credit in the macroeconomy." *Federal Reserve Bank of New York Quarterly Review*, 18, 1992–93, pp. 50–70.

Bernanke, B.S. and Blinder, A.S. "Credit, money, and aggregate demand". *American Economic Review Papers and Proceedings*, 78, 1988, pp. 435–39.

Bernanke, B.S. and Blinder, A.S. "The federal funds rate and the channels of monetary transmission."

Amercian Economic Review, 82, 1992, pp. 901–21.

Bernanke, B.S. and Gertler, M. "Agency costs, net worth, and business fluctuations." *American Economic Review*, 79, 1989, pp. 14–31.

Bernanke, B.S., Gertler, M. and Gilchrist, S. "The financial accelerator in a quantitative business cycle framework." In J.B. Taylor and M. Woodford, eds, *Handbook of Macroeconomics*, vol. 1C. Amsterdam: Elsevier North-Holland, 1999, pp. 1341–93.

Chari, V.V. and Kehoe, P.J. "On the robustness of herds." *Federal Reserve Bank of Minneapolis Working Paper* no. 622, 2002.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a new Keynesian perspective."

Journal of Economic Literature, 37, 1999, pp. 1661–707.

Driscoll, J.C. "Does bank lending affect output? Evidence from the U.S. states." *Journal of Monetary Economics*, 51, 2004, pp. 451–71.

Fama, E. "Banking in the theory of finance." *Journal of Monetary Economics*, 27, 1980, pp. 39–57.

Gertler, M. and Gilchrist, S. "Monetary policy, business cycles and the behavior of small manufacturing firms." *Quarterly Journal of Economics*, 109, 1994, pp. 309–40.

Guariglia, A. and Mateut, S. "Credit channel, trade credit channel, and inventory investment: evidence from a panel of UK firms." *Journal of Banking and Finance*, 30, 2006, pp. 2835–56.

Hubbard, R.G. "Is there a credit channel for monetary policy?" *Federal Reserve Bank of St. Louis Review*, 77, 1995, pp. 63–77.

Jaffee, D. and Russell, T. "Imperfect information, uncertainty and credit rationing." *Quarterly Journal of Economics*, 90, Nov. 1976, pp. 651–66.

Jaffee, D. and Stiglitz, J.E. "Credit rationing." In B. Friedman and F. Hahn, eds, *The Handbook of Monetary Economics*, Vol. II. Amsterdam: North-Holland, 1990, pp. 837–88.

Kashyap, A.K. and Stein, J.C. "Monetary policy and bank lending." In N.G. Mankiw, ed., *Monetary Policy*. Chicago: University of Chicago Press, 1994, pp. 221–56.

Kashyap, A.K. and Stein, J.C. "The role of banks in monetary policy: A survey with implications for the European monetary union." *FRB of Chicago Economic Perspectives*, 21, 1997.

Kashyap, A.K. and Stein, J.C. "What do one million observations on banks have to say about the transmission of monetary policy?" *American Economic Review*, 90, 2000, pp. 407–28.

Kashyap, A.K., Stein, J.C. and Wilcox, D.W. "Monetary policy and credit conditions: evidence from the composition of external finance." *American Economic Review*, 83, 1993, pp. 78–98.

Keynes, J.M. "Alternative theories of the rate of interest." *Economic Journal*, 47, 1937, pp. 241–52.

King, S.R. "Monetary transmission: through bank lending or bank liabilities." *Journal of Money, Credit and Banking*, 18, 1986, pp. 290–303.

Kiyotaki, N. and Moore, J. "Credit cycles." *Journal of Political Economy*, 105, 1997, pp. 211–48.

Lang, W.W. and Nakamura, L.I. " 'Flight to quality' in bank lending and economic activity." *Journal of Monetary Economics*, 36, 1995, pp. 145–64.

Ludvigson, S. "The channel of monetary transmission to demand: evidence from the market for automobile credit." *Journal of Money, Credit and Banking,* 30, 1998, pp. 365–83.

Minsky, H.P. *Stabilising an Unstable Economy*. New Haven, CT: Yale University Press, 1986.

Modigliani, F. and Miller, M.H. "The cost of capital, corporate finance, and the theory of investment."

American Economic Review, 48, 1958, pp. 261–97.

Oliner, S.D. and Rudebusch, G.D. "Is there a broad credit channel for monetary policy?" *Federal Reserve Bank of San Francisco Economic Review*, 1, 1996, pp. 300–09.

Ramey, V. "How important is the credit channel in the transmission of monetary policy?" *Carnegie-Rochester Conference Series on Public Policy*, 39, 1993, pp. 1–45.

Romer, C.D. and Romer, D.H. “New evidence on the monetary transmission mechanism.” *Brookings Papers on Economic Activity*, 1, 1990, pp. 149–98.

Stiglitz, J. and Weiss, A. “Credit rationing in models with imperfect information.” *American Economic Review*, 71, 1981, pp. 393–410.

Walsh, C.E. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press, 2003.

Williamson, S.D. “Costly monitoring, loan contracts and equilibrium credit rationing.” *Quarterly Journal of Economics*, 102, 1987, pp. 135–45.

17 Macro models and perspectives on the neutrality of money

This chapter continues the discussion of the effectiveness of monetary policy. It starts with the compact Lucas–Sargent–Wallace model, which is a popular platform for the modern classical approach. It then examines popular Keynesian and new Keynesian compact models. The emphasis of this chapter is on the presentations of compact, testable models and their findings.

Economists' and central bankers' thinking on the relevance and impact of monetary policy is presented in the conclusions.

Key concepts introduced in this chapter

- Monetary policy ineffectiveness proposition
- Neutrality of money
- Lucas–Sargent–Wallace model
- Lucas critique of estimated equations
- Keynesian supply rule
- New Keynesian Taylor rule
- New Keynesian IS equation
- New Keynesian model
- Hysteresis

The use of compact models to examine the impact of monetary policy on the macroeconomy is common in the macroeconomic literature. This chapter examines several of these models. For the modern classical approach, we have selected the Sargent and Wallace (1976) model. For the Keynesian and new Keynesian approaches, we have selected the models of Gali (1992), Clarida *et al*. (1999) and Levin *et al*. (2001) for the closed economy, and that of Ball (1999, 2000) for the open economy.

Sections 17.1 and 17.2 analyze the effects of systematic and unanticipated money supply changes on output and prices in the context of the Lucas–Sargent–Wallace model. Section 17.3 shows the validity of the Lucas critique in this model. Sections 17.4 to 17.7 cover the empirical evidence on the issue of monetary neutrality. Sections 17.8 and 17.9 present compact forms of the new Keynesian models and examine their validity. Section 17.10 sums up the empirical evidence on money neutrality and Section 17.11 suggests getting away from dogma. Section 17.12 provides a brief comment on hysteresis in the overall context of long-run money neutrality.

The conclusions of this chapter present a summing-up of our knowledge as reflected in the writings of some of the major protagonists on the neutrality debate.

The Lucas–Sargent–Wallace (LSW) analysis of the classical paradigm

The Lucas (1972, 1973) model presented in Chapter 14 serves as the underlying supply behavior of the modern classical approach. Based on this supply rule, the 1976 Sargent–Wallace model and its variants represent the most commonly used format for deriving the implications of the modern classical model and testing them. Since this model is based on Lucas (1972, 1973) and incorporates the Lucas supply function, we refer to it as the Lucas–Sargent–Wallace (LSW) model. It explicitly specifies the markets for commodities and money, with the assumption of equilibrium in these markets. However, the labor market is not explicitly modeled but is replaced, in conjunction with the production function, by the Friedman–Lucas supply function.

For our presentation of this model, we specify the change in the supply of commodities by stating the expectational error-based Lucas supply function or rule[1] as:

$$\mathrm{D}y^{s}_{t} = \alpha \mathrm{D}y_{t-1} + \beta(p_t - p^{e}_t) + \mu_t \qquad 0 \le \alpha \le 1, \beta > 0 \tag{1}$$

where all lower-case variables are in *logs*, the superscripts s and d stand respectively for supply and demand, and:

y = output
y^{f} = full employment
output $\mathrm{D}y_t$ = output gap $(\mp_t\ y_t$
$- y^{\mathrm{f}})$

p = price level
p^{e} = expected price level, with expectations formed one period earlier
μ = random term.

Note that D*y* is the deviation from full-employment output and not the previous period's output. μ and the other random terms in this chapter have a zero expected value and are independent of the other variables in the model. Equation (1) differs from the Friedman–Lucas supply rule derived in Chapter 14 by including a lagged term in output, with $\alpha > 0$. This is done in order to capture the commonly observed serial correlation of output over time. This alteration can be explained by adducing adjustment costs for employment, so that the marginal product of labor depends on both the current and last period's output.[2] Note that (1) also differs from the Friedman–Lucas supply rule in that it allows an output gap to exist even when there are no errors in price expectations: current output can differ from its full-employment level if there was an output gap in the preceding period. However, the output gap gets eliminated at the rate α per period. (1) also allows the full-employment output to change between periods. As shown in Chapter 13, the dependence of y^{s}_{t} on $(p_t - p^{e}_t)$ can be justified through the expectations-augmented Phillips curve, with nominal wages either fixed for the duration of the labor contract or fully flexible but with imperfect information about the price level and with expectational errors in perceived relative prices.[3]

[1] We are subsuming the expectations-augmented Phillips curve in the Lucas supply rule in this chapter.

[2] Such a term is needed for classical models since, in its absence, the deviations of output from full employment would depend only upon errors in expectations. Since the current classical models also assume rational expectations in which the errors in expectations are random, (1) with rational expectations would imply that output variations over time would be only random, even though business cycles show serial correlation in output.

3 Lucas (1973) specifies that production takes place on isolated points called islands; the selling price on the island is known but prices on other islands are not known. The demand for labor is a function of the real

The demand for output is specified as:

$$y^d_t = \theta(m_t - p_t) + \eta_t \qquad \theta > 0 \tag{2}$$

where:

y^d = aggregate demand
m = nominal money supply
η = random term.

Again, all variables are in logs. (2) represents the aggregate demand function and is a reduced-form relationship derived from the IS–LM relationships. It ignores fiscal policy for two reasons. One is that the effects of monetary and fiscal expansions in the IS–LM models are similar, so that keeping only one policy variable simplifies the model. In this sense, fiscal policy does influence aggregate demand and its stance is proxied in (2) through m_t . The other reason is that the new classical model with Barro's Ricardian equivalence theorem, outlined in Chapter 13 above, implies that fiscal deficits do not change aggregate demand. Only increases in the money supply do so, with the result that (2) becomes the accurate representation of the aggregate demand function, irrespective of the fiscal policy stance.

Equation (2) differs from the aggregate demand function in Lucas's model in Chapter 14 by making explicit the role of the nominal money supply in the determination of aggregate demand. But it also necessitates the specification of the money supply function, which is done by the monetary policy rule:

$$m_t = m_0 + \gamma \mathrm{D}y_{t-1} + \xi_t \qquad \gamma < 0 \tag{3}$$

Equation (3) makes the plausible assumption that the monetary authority increases the money supply if D$y_{t\ \ 1}$ < 0, i.e. if output last period was below the full employment level.

The equilibrium condition for the commodity market is:

$$y_t = y^d_t = y^s_t \tag{4}$$

which also translates to:

$$\mathrm{D}y_t = \mathrm{D}y^d t = \mathrm{D}y^s\ t$$

The above model has to be supplemented by an expectations hypothesis. Assuming the rational expectations hypothesis (REH), we have:

$$p^e_t = \mathrm{E}p_t \tag{5}$$

where Ep_t represents the expected price conditional on information available at the beginning of period t and (p_t–Ep_t) is a random variable with zero mean. In particular, p_{t-1} is part of this information set.

wage rate in terms of the island price, which on average in the aggregate for all islands equals the actual price, while the supply of labor is a function of the expected real wage in terms of the expected price level over all islands.

The complete LSW model, labeled in this chapter *LSW Model I*, consists of equations (1) to (5). With all variables in logs, this model is:

$$Dy^s_t = \alpha Dy_{t-1} + \beta(p_t - p^e_t) + \mu_t \qquad \alpha, \beta > 0 \qquad (1)$$

$$y^d_t = \theta\,(m_t - p_t) + \eta_t \qquad \theta > 0 \qquad (2)$$

$$m_t = m_0 + \gamma\, Dy_{t-1} + \xi_t \qquad (3)$$

$$y_t = y^d_t = y^s_t \qquad (4)$$

$$p^e_t = Ep_t$$

The basic question we wish to investigate within this model is whether monetary policy can be manipulated to increase output. More explicitly, we want to investigate whether there are particular values of m_0 and γ in (3) which optimize y_t. To answer this, we need to derive the reduced-form equation for y_t . From (1), (4) and (5), and taking the expectation of y_t , we have:

$$\begin{aligned} EDy_t &= \alpha EDy_{t-1} + \beta[Ep_t - E(Ep_t\,)] + E\mu_t \\ &= \alpha Dy_{t-1} \end{aligned} \qquad (6)$$

Substituting (4) in (2) and taking its expectation, with $E\eta_t = 0$, gives:

$$\begin{aligned} Ey_t &= \theta\,(Em_t - Ep_t\,) + E\eta_t \qquad (7) \\ &= \theta\,(Em_t - Ep_t\,) \end{aligned}$$

Subtracting (7) from (2) yields:

$$y_t - Ey_t = \theta(m_t - Em_t) - \theta(p_t - Ep_t) + \eta_t \qquad (8)$$

Subtracting y^f_t from both sides, we get:

$$Dy_t = EDy_t + \theta(m_t - Em_t) - \theta(p_t - Ep_t) + \eta_t \qquad (8^J)$$

where, from (3),

$$Em_t = m_0 + \gamma\, Dy_{t-1} \qquad (9)$$

so that:

$$m_t - Em_t = \xi_t \qquad (10)$$

From (1), (4) and (5),

$$p_t - \mathrm{E}p_t = (1/\beta)(\mathrm{D}y_t - \alpha \mathrm{D}y_{t-1}) - (1/\beta)\mu_t \tag{11}$$

In (8^J), replacing the relevant terms on the right-hand side from (6), (10) and (11) gives:

$$\begin{aligned} Dy_t &= \alpha Dy_{t-1} + \frac{\theta\mu_t + \beta\theta\xi_t + \beta\eta_t}{\beta + \theta} \\ &= \alpha Dy_{t-1} + W_t \end{aligned} \tag{12}$$

where:

$$\psi_t = \frac{\theta\mu_t + \beta\theta\xi_t + \beta\eta_t}{\beta + \theta}$$

Hence, the current output gap is the correction by α of last period's gap plus a new disturbance. (12) represents the core implication of the LSW model for the determination of output.

The ineffectiveness proposition of the LSW model for monetary policy

Since neither of the systematic policy parameters m_0 and γ occur in (12), the authorities cannot use systematic monetary policy to change y_t . Since ξ_t is in (12), errors in predicting money supply do affect y_t , but if the authorities were to increase such errors, it would only increase the variance of y_t without increasing the output level and, therefore, without constituting a sensible policy. There is therefore no optimal monetary policy in this model. Even though output can differ from its full-employment level, non-random policy does not have any long-run or even short-run effects on real output; it can neither cause nor improve nor worsen booms or recessions. These results are similar to those derived from the Lucas model in Chapter 14. The implication of the futility of systematic money supply changes to alter output – and by implication, employment and unemployment – is known as the "the ineffectiveness of demand policies" or the "demand policy irrelevance" result. This is so even if current output is below the full-employment rate. Hence, there is no stabilization role for monetary policy in this model. Note that, just as systematic money supply changes can alter aggregate demand but not output and employment, systematic exogenous increases in investment, consumption, net exports, etc., also cannot change output and employment by virtue of their impact on aggregate demand. But their random components can do so.

The LSW model does not support Keynesian policy recommendations on the use of monetary policy to reduce deviations from full-employment output. However, this is understandable since the model is not a Keynesian one to start with. In particular, its supply equation (1) is the Friedman–Lucas supply rule, which embodies, under rational expectations, the neutrality of systematic changes in the money supply.

Price level in the LSW model

The reduced form for p_t for this model is obtained by substituting (12)^4 in (2). This gives:

$$p_t = m_t - \frac{1}{\theta} y_t^f - \frac{\alpha}{\theta} Dy_{t-1} - \frac{1}{\theta}\psi_t + \frac{1}{\theta}\eta_t \tag{13}$$

[4] To do so, first replace Dy_t by ($y_t - y_t^{\text{f}}$).

Equation (13) implies that $\partial p_t / \partial m_t = 1$. Hence, since both p and m are in logs, prices rise proportionately with the *overall* increase in the nominal money supply, whether it is due to the systematic factors γ and m_0 or the random component ξ_t .

To find p^e_t, since $p^e_t = E(p_t)$, taking the rational expectation of (13) implies that:

$$p^e_t = E(m_t) - (1/\theta)y^f_t - (\alpha/\theta)Dy_{t-} \quad (14)$$

From (14) and (9),

$$p^e_t = m_0 - (1/\theta)y^f_t + \{\gamma - (\alpha/\theta)\}Dy_{t-} \quad (15)$$

In (14) and (15), $\partial p^e_t/\partial E m_t = 1$, so that the expected price level rises proportionately in the same period with the *systematic* component of the nominal money supply, and responds to the policy parameters γ and m_0, but not to the random part ξ_t of the money supply. Hence, increases in systematic money supply (and systematic demand) proportionately change both the price level *and* the expected price level but do not cause a change in output, whereas random changes in the money supply change the price level and output but not the expected price level.

Diagrammatic analysis of output and price level in the LSW model

The effect of a systematic monetary increase in the LSW model is shown in Figure 17.1. From (2), the aggregate demand curve AD has a negative slope and shifts to the right from AD$_0$ to AD$_1$ with a monetary expansion. From (1), the (short-run) aggregate supply curve SAS has a positive slope and shifts to the left from SAS$_0$ to SAS$_1$ with an increase in the *expected* price level. Since the latter, from (14), increases proportionately with a systematic monetary expansion, such a monetary expansion results in proportionate shifts of both curves, and the economy goes directly from point a to point c without an intermediate increase in output. However, a random increase in the money supply cannot be anticipated, so that the expected price level does not increase as a result and the supply curve does not shift from SAS. But the demand curve does go from AD$_0$ to AD$_1$, with the new equilibrium at point b causing an increase in both prices and output. Hence, while the systematic increases in money supply do not bring about an increase in output, unexpected changes in it do so.[5]

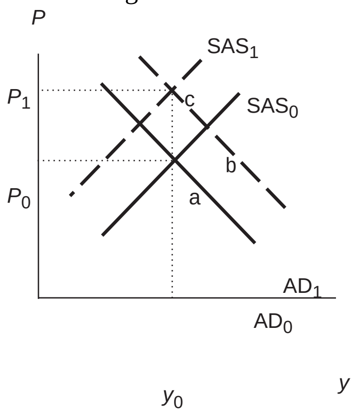


Figure 17.1

5 If we continue the story further, the unexpected increase in the current price level will have become anticipated in the following period, so that, barring new sources of shifts in aggregate demand and supply, the economy will

A compact (Model II) form of the LSW model

Since the price level is a function of the money supply and the expected price level is a function of the expected money supply, the above model with rational expectations is often replaced by the following more compact one, labeled by us as the LSW Model II.

Lucas supply rule:

$$Dy_t = \alpha Dy_{t-1} + \beta(m_t - m^e_t) + \mu_t \qquad \alpha, \beta > 0 \qquad (16)$$

Money supply rule:

$$\gamma < 0 \qquad (17)$$

$$m_t = m_0 + \gamma Dy_{t-1} + \xi_t$$

$$m^e_t = Em_t \qquad (18)$$

Note that the assumption of equilibrium in the commodity market has already been incorporated in (16). From (17) and (18),

$$m_t - m^e_t = \xi_t \qquad (19)$$

Hence, from (16) and (19),

$$Dy_t = \alpha Dy_{t-1} + \beta \xi_t + \mu_t \qquad (20)$$

where Dy_t is again independent of the systematic monetary policy parameters m_0 and γ, so that changes in these parameters will not change y_t. Hence, the policy invariance result also holds in this model. Note that this is a strong and seemingly surprising result: monetary policy cannot correct for the output gap that arises because of persistence in the model. Even if $y_t < y^f_t$, an expansionary systematic monetary policy cannot reduce the output gap, nor increase it. Further, the sources of the output gap are persistence, random disturbances and fluctuations in the full-employment output, with actual output lagging behind the full-employment level. These provide the basis for the real business cycle (RBC) theories. Given that $0 < \alpha < 1$, the deviations of output from its full-employment level are self-correcting over time.

For another – though misleading – pattern of derivation in the above model, we have from (17):

$$Em_t = m_0 + \gamma Dy_{t-1} \qquad (21)$$

be at point c after the current period. Therefore, the increase in output is likely to be short lived. Its duration will depend upon the time it takes the public to correct for erroneous price expectations.

so that, from (16), (18) and (21):

$$\begin{aligned} \mathrm{D}y_t &= \alpha \mathrm{D}y_{t-1} + \beta m_t - \beta(m_0 + \gamma\, \mathrm{D}y_{t-1}) + \mu_t \\ &= (\alpha - \beta\gamma)\mathrm{D}y_{t-1} - \beta m_0 + \beta m_t + \mu_t \end{aligned} \tag{22}$$

In (22), Dy_t depends upon the policy parameters m_0 and γ, so that we get the impression that output will depend upon systematic monetary change. However, this would be erroneous since substituting (17) in (22) to eliminate m_t again gives (20), which establishes the systematic policy irrelevance result.

The Lucas critique of estimated equations as a policy tool

Suppose the economic researcher were to use (20) to set up an estimating equation of the form:

$$y_t = a_0 + a_1 OY_t + \xi_t \qquad a_1 \geq 0 \tag{23}$$

where y is output, Y is nominal aggregate demand and ξ is white noise, and found the estimated value of a_1 to be positive (i.e. $\hat{a}_1 > 0$). As argued above, if the policy maker increased aggregate demand by more than was experienced during the estimation period, a_1 would shift, so that $\hat{a}_1$ would no longer be the relevant magnitude under the revised demand policy. Hence, a maintained shift in the expansion of demand does not leave the estimated parameters constant. This is known as the *Lucas critique* of estimated functions of the Lucas–Phillips curve type; if the underlying model of the economy is as set out by Lucas, the parameters of functions such as (23) are not invariant with respect to policy shifts.

Lucas (1973) estimated a variant of (23) for a cross section of countries and found that countries with low rates of inflation (such as the USA in the 1950s and 1960s) showed evidence of a positive relationship between output and demand increases while those with hyper-inflation (such as Argentina in the 1950s and 1960s) did not. He thereby concluded that this relationship shifts as inflation increases, and that the Phillips curve tradeoff does not hold at high and persistent rates of inflation.

Similarly, an estimating equation with unemployment u as the dependent variable would be:

$$u_t = b_0 + b_1 OY_t + \xi_t \qquad b_1 \leq 0 \tag{24}$$

Note that an estimated function such as (24) does not distinguish between the expectations-augmented Phillips curve, based on errors in price level expectations in labor markets, and the Lucas supply rule, based on errors in relative commodity prices, so that the estimated coefficients would reflect the influence of both types of errors. Equations such as (23) with output as the dependent variable or (24) with unemployment as the dependent variable have been estimated for a variety of countries and for a variety of time periods. Lucas's conclusion, that such functions are often unstable under demand policy shifts, is now well established.[6]

[6] Another aspect of the preceding discussion and of the Lucas critique is that economic agents learn from experience, so that their expectations shift if policy shifts. This brings us back to the question of the expectations hypotheses and the role of learning in them. A number of learning mechanisms have been proposed and the speed at which expectational errors are eliminated is derived for them. However, they go

beyond the concerns of this book and are not examined herein.

However, another of Lucas's conclusions was that systematic demand increases do not change real output and unemployment. This has not always been supported in empirical studies, as the stylized facts on money and output at the beginning of Chapter 14 show.

The Lucas critique in the LSW model

To check on the applicability of the Lucas critique in the LSW model comprising equations (16) to (18), restate (22) compactly as:

$$y_t = a_0 + a_1 y_{t-1} + a_2 m_t + \mu_t \tag{25}$$

where a_0 $y^f{}_t$ $(\alpha$ $\beta\gamma$ $)y^f{}_{t-1}$ βm_0, a_1 $(\alpha$ $\beta\gamma$ $)$ and a_2 β. If (25) is estimated, it would yield the estimated values of a_1 and a_2 as $\hat{a}_1$ and $\hat{a}_2$. If $\hat{a}_2 > 0$, it would be tempting to conclude that the authorities could increase the money supply to increase output. However, as we have shown earlier in Section 17.2, this would not be a valid conclusion. A policy change in the money supply rule would mean a shift in the values of m_0 or/and γ . But, as these shift, a_1 and a_2 in (25) would shift, as can be seen from a glance at (22), of which (25) is only the compact form. Hence, it cannot be assumed that a_1 and a_2 are invariant to a policy change. The Lucas critique therefore applies to (25), so that this equation cannot be used as a tradeoff for policy formulation. Further, (25) cannot be used for regression estimates across policy regimes since such estimation assumes constant parameters.[7]

Testing the effectiveness of monetary policy: estimates based on the Lucas and Friedman supply models

The Friedman–Lucas supply function can be stated with output, employment, unemployment or another real variable, such as the real rate of interest as the dependent variable, and with the expectational errors in absolute prices, in aggregate demand or with just money supply as the independent variable. From a monetary policy perspective, the form of the Friedman–Lucas supply function that is usually tested is:

$$y_t = a_0 + a_1(M_t - M^e{}_t) + \Sigma_j a_j z_{jt} + \mu_t \qquad a_1 > 0 \tag{26}$$

where:

M = nominal money supply
M^e = expected nominal money supply
z_j = other exogenous variables
μ = random term.

Equation (26) focuses on money supply as the sole policy variable determining aggregate demand. Under rational expectations,

$$M^e_t = EM_t \tag{27}$$

[7] For consistent and unbiased estimates using data from a given policy regime – i.e. with constant true values of m_0 and γ – estimate (21) and (25). Since there are five reduced form coefficients (a_0, a_1, a_2, m_0 and γ) in equations
(24) and (26), while there are only four structural ones (m_0, γ , α, β) in (16) to (18), cross-equation restrictions

implied by (25) will have to be imposed and a simultaneous estimation procedure used.

where EM_t is proxied by its estimated value E$M_t = \hat{M}_t$. Similarly, if needed, the estimated forms of the other variables in (26) can be used to modify this equation, but are omitted here. Therefore, the estimated form of (26) becomes:

$$y_t = a_0 + a_1(M_t - \hat{M}_t) + \Sigma_j a'_j z_{jt} + \mu_t \qquad a_1 > 0 \tag{28}$$

Equation (28) is a commonly used form of the modern classical output hypothesis for the short run. In the long run, $M = \hat{M}$, so that money supply changes cannot affect y.

A procedure for segmenting the money supply changes into their anticipated and unanticipated components

To estimate (28), we need the value of M_t^e. Its value is usually the one specified by the REH and is a function of the information available to the economic agent. Assuming that the central bank controls the national money supply, the relevant knowledge would be that of the public on central bank behavior and on the policy rule that the central bank follows. Assume that this is a rule that gives the money supply function as:

$$M_t = \Sigma_i \alpha_i x_{it} \tag{29}$$

where x_t is a set of exogenous and predetermined variables. Adding in a disturbance term η gives the money supply rule function:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{30}$$

Under the REH, the public is assumed to know the policy function (29) and use the estimated values $\hat{\alpha}_i$ from (30) to calculate the estimated value $\hat{M}_t$, where $\hat{M}_t$ is the rational expectations' proxy for the anticipated money supply, so that:

$$\hat{\eta}_t = M_t - \hat{M}_t \tag{31}$$

η_t is the proxy, under the rational expectations hypothesis, for the unanticipated money supply.

The nested form of the Lucas model

The nested form of (28) and (26), incorporating both $\hat{\eta}_t$ and $\hat{M}$, specifies the linear (or log-linear) estimating equation as:

$$y_t = \beta_0 + \beta_1\hat{M}_t + \beta_2\hat{\eta}_t + \Sigma_j \gamma_j z_{jt} + \mu_t \tag{32}$$

If $\hat{\beta}_1 = 0$, the anticipated values of money supply do not affect real output, so that this finding would be consistent with the modern classical hypothesis; but if $\hat{\beta}_1 > 0$, the modern classical hypothesis is rejected.[8,9]

Barro's test of the Lucas model: a joint test of neutrality and rational expectations

Barro (1977), in one of the earliest articles applying the REH to the Friedman–Lucas supply rule, used a two-step OLS procedure to test jointly for rational expectations and neutrality. In the first stage, he estimated by OLS a forecasting equation for the money supply. Under the assumption of rational expectations, the calculated value of the money supply from this estimation was used as the proxy for its anticipated value and the residual was used as the unanticipated value. The impact of these on the dependent real variable – in his case, unemployment – was then estimated in a second equation. Using this procedure, Barro (1977) reported the following estimated functions, using annual data for the USA for 1946–73:

$$\begin{aligned}\ln[U/(1-U)]_t = {} & -3.07 - 5.8\mathrm{DMR}_t - 12.1\mathrm{DMR}_{t-1} - 4.2\mathrm{DMR}_{t-2} - 4.7\mathrm{MIL}_t \\ & + 0.95\mathrm{MINW}_t \end{aligned} \tag{33}$$

$$\begin{aligned}\mathrm{D}\hat{M}_t = {} & 0.087 + 0.24\mathrm{DM}_{t-1} + 0.35\mathrm{DM}_{t-2} + 0.082 \ln \mathrm{FEDV}_t \\ & + 0.027 \ln [U/(1-U)]_{t-1} \end{aligned} \tag{34}$$

where:

$\mathrm{D}M$ = growth rate of money supply
$\mathrm{D}\hat{M}$ = estimated growth rate of money supply
DMR = unanticipated money growth rate ($=\mathrm{DM}-\mathrm{D}\hat{M}_t$)
MIL = military size
MINW = minimum wage
FEDV = federal government expenditures relative to their normal level
U = unemployment rate.

Equation (33) is a version of the Friedman–Lucas supply rule and (34) is a money supply rule. Barro's estimates showed that unanticipated money growth was significant in explaining

[8] $\beta_1 < 0$ indicates that increases in the money supply are detrimental for output in the economy, as when they increase the degree of uncertainty and reduce investment or lead otherwise to a diversion of resources to less efficient uses in the economy.

[9] A simplified form of this system can give erroneous results, as Mishkin (1982) showed. Consider the following simple system:

$$y_t = a_1\hat{M}_t + \Sigma_j g_j z_{jt} + \theta_t \tag{1}$$

$$M_t = b_1 y_{t-1} + W_t \tag{2}$$

which imply that:

$$y_t = a_1 b_1 y_{t-1} + \Sigma_j g_j z_{jt} + \theta_t \tag{3}$$

where $\hat{M}_t$ does not occur as an explanatory variable in (3), so that it would appear that (3) rejects the Keynesian hypothesis when in fact this hypothesis was part of the initial model in the form of (1). This problematic result arose because (2) involved only the lagged terms of y as explanatory variables, so that (1) and (2) were not identified through the reduced form (3), thereby making it impossible to judge from (3) whether the anticipated part of the money supply affected real output or not. Hence, estimating (3) does not allow us to discriminate between the two hypotheses.

current unemployment. They also showed that when the total money supply, current and with two lags, replaced the unanticipated money supply terms in (33), their coefficients were not significant. Barro concluded that his study supported the modern classical hypothesis based on Lucas (1972, 1973) and not the Keynesian one.

Separating neutrality from rational expectations: Mishkin's test of the Lucas model

Mishkin (1982) objected to Barro's estimation procedure since it only provided a test of the joint hypotheses of the neutrality of money and rational expectations, without providing separate results on each of these hypotheses.[10] To understand Mishkin's objection, note that Barro's estimation system was of the form:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{35}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha_i x_{it-j}) + \mu_t \tag{36}$$

where x_{it} were a set of exogenous or predetermined variables for determining the money supply. The output equation (36) embodies both neutrality and rational expectations. It also allows lags in the impact of the unanticipated money supply. Determinants of output other than the money supply have been left out of this equation, for simplification. The money supply equation (35) is essentially the same as (30), while (36) has been obtained by substituting the estimate of M_t from (30) in (28). This system imposes rational expectations since α_i in the money equation (35) also appears in the output equation (36). The neutrality property is imposed in (36) since the coefficients on EM_t are a priori set at zero.

To test for rational expectations and neutrality separately, the estimation system should be:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{37}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha^*_i x_{it-j}) + \sum_{j=0}^{n} \gamma_j \alpha^*_{i,t-j} x_{i,t-j} + \mu_t \tag{38}$$

where (38) is the nested equation (32). Rational expectations require $\alpha^*_i = \alpha_i$, while neutrality requires $\gamma_j = 0$.

Therefore, maintaining the rational expectations hypothesis – that is, setting $\alpha_i = \alpha^*_i$ – while relaxing the neutrality assumption implies testing the system:

$$M_t = \Sigma_i \alpha^*_i x_{it} + \eta_t \tag{39}$$

n n

$$y_t = \beta_0 + \overset{j=0}{-}\ \beta_j(M_{t-j} - \Sigma_i \alpha_i x_{it-j}) + \overset{j=0}{-}\ \gamma_j \alpha_{i,t-j} x_{i,t-j} + \mu_t \qquad (40)$$

[10] Mishkin also argued that while the two-step OLS estimation procedure will yield consistent parameter estimates, they do not generate valid F-test statistics, thereby resulting in inconsistent estimates of the standard errors of the parameters and test statistics, which do not follow the assumed F-distribution. He used the Full Information Maximum Likelihood (FIML) procedure for the nonlinear joint estimation for his systems.

The null hypothesis of neutrality – that is, γ_j_0 – can be tested by comparing the estimates of the system (39) and (40) with those from (35) and (36). Maintaining the neutrality hypothesis while testing for rational expectations requires estimating:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{41}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha^*_i x_{it-j}) + \mu_t \tag{42}$$

The null hypothesis of α^*_i α_i is tested by comparing the estimates from (41) and (42) against those from (35) and (36).

Mishkin's tests for the USA on quarterly data for 1954–76 used unemployment and output as the dependent variables, and nominal GNP and the rate of inflation among the independent variables. He reported that while the REH was not rejected by the data, neutrality was. Further, the estimated coefficients of the anticipated and unanticipated demand variables were very similar in magnitude. Therefore, Mishkin's results supported the Keynesian hypothesis and rejected the modern classical one on the key issue of the neutrality of anticipated aggregate demand and of demand management policies. These results did not reject the REH, which is merely a procedure for modeling expectations and, as noted earlier, is not a priori inconsistent with either the Keynesian or the modern classical theories.

Among several other studies, Frydman and Rappoport (1987) also tested for the distinction between the impact of anticipated and unanticipated monetary policy by examining the impact of the growth rate of money on output. Their findings reject this distinction for the short-run determination of output, with this rejection robust to different specifications of rational expectations and of the employment level of output.

We conclude that there is by now very substantial evidence that this distinction is not valid and, further, that anticipated monetary policy is not neutral, at least for the short run. These findings are not surprising in view of the stylized facts listed at the start of Chapter 14. Therefore, models of the Lucas and the Sargent and Wallace variety do not provide a useful basis for macroeconomic analysis and policy.

Distinguishing between the impact of positive and negative money supply shocks

It is sometimes argued that decreases in the money supply are likely to have a stronger impact than increases in it. There can be several reasons for this. Among these are:

1. A decrease in the money supply represents a decrease in credit in the economy, so that borrowers are forced to curtail their economic activities, and this reduces output in the economy. By comparison, an increase in the money supply means a greater willingness by the financial intermediaries to lend, which does not result in the same urgency to borrow as a decrease in loans to repay.[11]
2. Contractionary policies are likely to be pursued during booms with full employment and a high demand for investment and additional borrowing, whereas expansionary policies are likely to be pursued during recessions when firms generally face inadequate demand

[11] An analogy sometimes used is that one can pull on a string but not push on it.

for their products, often have excess capacity and do not have enough incentive to increase their investment and borrowing. That is, the impact of the two types of policies also depends on the phase of the business cycle with which the two are associated.

3. The economy is likely to possess some downward rigidity in prices and nominal wages whereas it possesses a higher degree of flexibility for increases in them. Decreases in the money supply run into this downward rigidity and are more likely to have real effects, while most or all of the impact of increases in the money supply could be only on prices and the nominal but not the real value of output.

The Barro and Mishkin tests can be modified to test for this differential impact. We illustrate this by modifying the Friedman–Lucas supply function to:

$$y_t = a_0 + a^{+}{}_1 M^{u}{}_t + \Sigma_j a_j z_{jt} + \mu_t \tag{43}$$

and its nested form (32) to:

$$y_t = a_0 + a^{+}{}_1 M^{u+}{}_t + a^{-}{}_1 M^{u-}{}_t + \beta^{+}{}_1 M^{e+}{}_t + \beta^{-}{}_1 M^{e-}{}_t + \Sigma_j \alpha_j z_{jt} + \mu_t \tag{44}$$

where:

M^{u+} = unanticipated increase in the money supply
M^{u-} = unanticipated decrease in the money supply
M^{e+} = expected (anticipated) increase in the money supply
M^{e-} = expected (anticipated) decrease in the money supply

and z_j are the other variables in the determination of output. The other aspects of the estimation procedures of Barro and Mishkin remain as specified earlier. In particular, the actual estimation should allow for lags and other independent variables. The Mishkin (1982) procedure specified above can be suitably modified to check on neutrality, rational expectations and asymmetrical effects.

Some empirical studies do report evidence of the asymmetrical effects – and non-neutrality – of monetary policy. Our earlier arguments suggest the possibility that negative shocks to money have greater impact than positive ones. For example, Cover (1992), using Mishkin's (1982) procedure, finds for US quarterly data that positive shocks to M1 had no effect on output, whereas negative ones did so. Ratti and Chu (1997) confirm for Japan the asymmetry between the effects of positive and negative shocks. They further report that unanticipated changes in a wide definition of money did not have a significant impact on output in Japan.

LSW model with a Taylor rule for the interest rate

As Chapter 13 pointed out, in recent decades many central banks have chosen to set the interest rate rather than the money supply. Assume that the central bank uses a contemporaneous Taylor rule with price level targeting[12] of the form:

$$r^{T}_t = r_0 + \lambda_y (y_t - y^{f}) + \lambda_P (P_t - P^{T}) \qquad \lambda_y, \lambda_P > 0 \tag{45}$$

There are two ways of integrating this rule in the LSW framework. One is to use the money demand function to derive the endogenous money supply that keeps the financial markets in

[12] This has been done to avoid having the price level in some equations and the inflation rate in others within

our model. *P* is the log of the price level and T indicates its target value.

equilibrium (see Chapter 13) and then use the resulting money supply function in the rest of the LSW model. The other method is to replace the money supply function in the original LSW model by the Taylor rule and make appropriate changes in the aggregate demand equation. The following derivations are based on the second method.

For the linear forms of the various functions, the IS equation for the open economy derived in Chapter 13 was:

$$y^d = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho}\rho^r\} + (1/\rho^r).\{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^r\}] \qquad (46)$$

where:

$$\alpha = \left[\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)}\right] > 0$$

and $\rho^r = \rho P/P^F$. The definitions of the symbols are as given in Chapter 13. For simplification, assume the general form of the IS equation to be:

$$y_t^d = y_0 - \theta_1 r^T - \theta_2 P_t \qquad (47)$$

The resulting LSW model with the stochastic forms of the preceding IS equation and the Taylor rule and all variables in logs is:

$$\mathrm{D}y^s_t = \alpha \mathrm{D}y_{t-1} + \beta(P_t - P^e_t) + \mu_t \qquad \alpha, \beta > 0 \qquad (48)$$

$$y^d_t = \theta_0 - \theta_1 r_t - \theta_2 P_t + v_t \qquad \theta_1, \theta_2 > 0 \qquad (49)$$

$$r_t = r^T_t + \eta_{1t} \qquad (50)$$

$$r^T_t = r_0 + \lambda_y \mathrm{D}y_t + \lambda_P (P_t - P^T) + \eta_{2t} \qquad \lambda_y, \lambda_P > 0 \qquad (51)$$

$$y_t = y^d_t = y^s_t \qquad (52)$$

$$P^e_t = \mathrm{E}P_t \qquad (53)$$

where $\mathrm{D}y = y_t - y^f$, μ is supply shocks, v is IS shocks, η_2 is the monetary policy shock and η_1 is the stochastic slippage in the control of the economy's interest rate by the central bank. All disturbances are taken to be white noise.

Taking the rational expectation of the aggregate supply equation (48), we have:

$$\mathrm{E}\mathrm{D}y^s_t = \alpha \mathrm{E}\mathrm{D}y_{t-1} + \beta[\mathrm{E}P_t - \mathrm{E}(\mathrm{E}P_t)] + \mathrm{E}\mu_t$$

so that:

$$= \alpha \mathrm{D}y_{t-1}$$

(54)

$$\mathrm{D}y^{s}_{t} - \mathrm{ED}y^{s}_{t} = \beta(P_{t} - \mathrm{E}P_{t}) + \mu_{t} \qquad (55)$$

Further, from (48) and (53):

$$P_t - EP_t = (1/\beta)(Dy_t - \alpha Dy_{t-1}) - (1/\beta)\mu_t \qquad (56)$$

Now, taking the rational expectation of the IS equation (49), we have:

$$Ey^d_t = \theta_0 - \theta_1 Er_t - \theta_2 EP_t \qquad (57)$$

Subtracting (57) from (49) yields:

$$y^d_t - Ey^d_t = -\theta_1(r_t - Er_t) - \theta_2(P_t - EP_t) + v_t \qquad (58)$$

where r_t is given by:

$$r_t = r_0 + \lambda_y Dy_t + \lambda_P(P_t - P^T) + \eta_{2t} + \eta_{1t} \qquad (59)$$

so that:

$$Er_t = r_0 + \lambda_y EDy_t + \lambda_P(EP_t - P^T) \qquad (60)$$

Hence:

$$r_t - Er_t = \lambda_y(Dy_t - EDy_t) + \lambda_P(P_t - Ep_t) + \eta_{2t} + \eta_{1t} \qquad (61)$$

Substituting (61) in (58) and imposing equilibrium, so that $y_t = y^d_t = y^s_t$, and $Dy_t = Dy^d_t = Dy^s_t$, we have:

$$y_t - Ey_t = -\theta_1\{\lambda_y(Dy_t - EDy_t) + \lambda_P(P_t - EP_t) + \eta_t\} - \theta_2(P_t - EP_t) + v_t \qquad (62)$$

where $\eta_t = \eta_{1t} + \eta_{2t}$. Since $y_t - Ey_t = Dy_t - EDy_t$, where $Dy = y - y^f$, (62) becomes:

$$Dy_t = EDy_t - \theta_1\lambda_y Dy_t + \theta_1\lambda_y EDy_t - \theta_1\lambda_P(P_t - EP_t) - \theta_1\eta_t - \theta_2(P_t - EP_t) + v_t \qquad (63)$$

$$(1 + \theta_1\lambda_y)Dy_t = (1 + \theta_1\lambda_y)EDy_t - (\theta_1\lambda_P + \theta_2)(P_t - EP_t) - \theta_1\eta_t + v_t$$

$$= [\{(1 + \theta_1\lambda_y)\alpha Dy_{t-1} - (1/\beta)(\theta_1\lambda_P + \theta_2)\{(Dy_t - \alpha Dy_{t-1}) - \mu_t\} - \theta_1\eta_t + v_t$$

$$\{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}Dy_t$$

$$= \{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}\alpha Dy_{t-1} - \{(1/\beta)(\theta_1\lambda_P + \theta_2)\}\mu_t - \theta_1\eta_t + v_t \qquad (64)$$

Replacing $\{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}$ by a_1 and $\{(1/\beta)(\theta_1\lambda_P + \theta_2)\}$ by a_2, we get:

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} - (a_2/a_1)\mu_t - (1/a_1)\theta_1\eta_t + (1/a_1)v_t \tag{65}$$

where α is the fraction by which the economy, on its own, adjusts the current period's output gap as a fraction of last period's gap. It is independent of the systematic policy parameters λ_y, λ_P embodied in the Taylor rule. However, random IS shocks (μ), monetary sector shocks (η_1) and policy shocks (η_2) do impact on the output gap, with this impact a function of various parameters, including those of policy. This conclusion is similar to that derived earlier from the LSW model with a money supply function. In particular, systematic policy parameters do not affect output in both models. This is especially surprising in the current model with the Taylor rule, which explicitly targets the output gap. This finding highlights the point that, in this model, the driving conclusions on output are given by the nature of the supply rule and the expectations hypothesis, not by the monetary policy rule, which determines systematic monetary policy. In the LSW model, this supply rule is the Lucas rule. Replacing it by a Keynesian or NK supply function is essential for showing the effectiveness of systematic policies in changing the output gap.

Note that the policy parameters do affect the impact on output of each of the disturbances.

Testing the effectiveness of monetary policy: estimates from Keynesian models

Using the LSW model with a Keynesian supply equation

A Keynesian supply function would differ from the Lucas one in several ways. One, it would allow both anticipated and unanticipated money supply changes to affect output. Two, as argued in Chapter 15, the effect of money supply changes on output would depend on the state of the economy and the degree of involuntary unemployment. Three, in Keynesian models, the impact of money supply changes on output can occur without a prior change in the price level, so that the LSW(II) model with the money supply as an explanatory variable is preferable to the LSW model with the price level. Further, note that, in the Keynesian context, the dependence of y_t on $y_{t\ 1}$ occurs because Keynesian models allow for staggered wage contracts longer than one period as well as the gradual adjustment of output to shocks.

Adapting the Friedman–Lucas supply function to a Keynesian format by replacing the unanticipated money supply $(m_t - m^e{}_t)$ by the total money supply m_t gives:

$$\mathrm{D}y_t = \alpha y_{t-1} + \beta_t m_t + \mu_t \qquad \alpha,\ \beta > 0 \tag{66}$$[13]

where the definitions of the symbols are as given earlier. As a reminder, note that the lower-case letters are logs of variables and Dy is the deviation from the full-employment level and not a change from the previous period's level. For Keynesian models, since the money multiplier depends on the state of the economy, the value of β_t should be a function of the output gap.

If we specify the complete model as consisting of the money supply function (17) and the Keynesian output supply function (66), we find that:

$$\mathrm{D}y_t = (\alpha + \beta_t \gamma)\mathrm{D}y_{t-1} + \beta_t m_0 + \mu_t + \beta_t \xi_t \tag{67}$$

where Dy_t depends upon the policy parameters m_0 and γ, which can be used to achieve the desired objectives with respect to y_t, and upon the money multiplier β_t, as well as on the previous period's deviation from full-employment output.

Equation (67) is not an appealing format for the Keynesian supply function. A better format seems to be one where the deviation of output from its full-employment level depends not

[13] β in true Keynesian models would properly not be a constant but would be a function of the deviation of output from full-employment level. However, Sargent and Wallace (1976) proposed $\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} + \beta m_t + \mu_t$ as the Keynesian reduced-form equation, with a constant β, and it is frequently cited as such.

on the money supply but on its change. The output supply function consistent with this idea is:

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} + \beta_t(m_t - m_{t-1}) + \mu_t, \qquad \alpha,\ \beta > 0 \tag{68}$$

The LSW money supply rule (17) can also be reformulated as:

$$m_t - m_{t-1} = \gamma_1(y_{t-1} - y_{t-2}) - \gamma_2(y^{\mathrm{f}}_{t-1} - y^{\mathrm{f}}_{t-2}) + (\xi_t - \xi_{t-1}) \qquad \gamma_1,\ \gamma_2 < 0 \tag{69}$$

This modification of the LSW money supply rule allows for differential response of the money supply to changes in actual output versus changes in full-employment output. Since γ_1, $\gamma_2 < 0$, the monetary authority decreases the money supply if the actual output rose last period and increases it if there was a rise in the full-employment output.

Equations (68) and (69) yield:

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} + \beta_t\gamma_1(y_{t-1} - y_{t-2}) - \beta_t\gamma_2(y^{\mathrm{f}}_{t-1} - y^{\mathrm{f}}_{t-2}) + \mu_t + \beta_t(\xi_t - \xi_{t-1}) \tag{70}$$

which implies that the monetary authority can change its policy parameters γ_1, γ_2 to affect output deviations from full employment. Therefore, policy irrelevance does not occur in this Keynesian model, irrespective of any assumption on the rationality of expectations. However, the Lucas critique will still apply since the coefficient of y_{t-1} in (70) depends on the monetary policy parameters γ_1 and γ_2, which will change with a policy shift.

Gali's version of the Keynesian model with an exogenous money supply

Chapter 15 reported the findings from Mishkin's (1982) test from reduced-form equations that money was not neutral while rational expectations were valid.

Gali (1992) uses a structural model that is Keynesian with a Phillips curve. His model, in logs, is:

IS equation:

$$y_t = \mu^{\mathrm{s}}_t + \alpha - \sigma\,(R_t - EOp_{t-1}) + \mu^{\mathrm{IS}}_t \tag{71}$$

LM equation:

$$m_t - p_t = \varphi y_t - \lambda R_t + \mu^{\mathrm{md}}_t \tag{72}$$

Money supply process:

$$Om_t = \mu^{\mathrm{ms}}_t \tag{73}$$

Phillips curve:

$$Op_t = Op_{t-1} + \beta(y_t - \mu^{\mathrm{s}}_t) \tag{74}$$

where the symbols designate the logs of the relevant variables, except for *R* (which stands for the level of the nominal interest rate). μ^{s}, μ^{IS}, μ^{md} and μ^{ms} are the stochastic processes for output supply, expenditures, money demand and money supply respectively. The IS and LM equations are consistent with the Lucas model, except that they provide greater detail.

The major difference between this Keynesian model and the LSW models lies in the specification of output supply. Gali specifies it by the Phillips curve, so that changes in the inflation rate between periods determine the variations μ^s in output from its equilibrium level, whereas the LSW model uses relative price misperceptions to explain such deviations. In the Gali model, output can change due to supply shocks through μ^s, or demand shocks due to μ^{IS}, μ^{md} or μ^{ms}. Positive demand shocks increase both output and prices while positive supply shocks increase output and decrease prices. Monetary shocks are transmitted to the real sector only through changes in the interest rate.

The segregation of the experienced shocks into four different types requires special assumptions on their origin or impact. Gali separated the supply shocks from the demand shocks by the assumption that the former have long-run effects on output while the latter do not. The IS shocks were separated from the money market shocks by the assumption that the latter do not have contemporaneous impact on aggregate demand in the same quarter, since their impact occurs through the changes in interest rates impinging on investment. The money demand and supply shocks were separated under three alternative assumptions, which we do not report here.

Gali's data was quarterly for the USA for 1955:I to 1987:IV. The monetary aggregate used was M1 and the interest rate was represented by the three-month Treasury bill rate. The findings supported the Keynesian claim that demand shocks do cause output changes while rejecting its claim that they, rather than supply shocks, were the dominant source of output fluctuations. Supply shocks had a substantial deflationary impact and accounted for about 70 percent of output variability over the business cycle. However, their impact on the nominal interest rate was small.

Increases in M1 increased output, reaching a peak in about four quarters, accompanied by increases in inflation and nominal interest rates but decreases in the real rate. While the money supply shocks accounted for most of the short-run variability of the nominal and real interest rates and for some variability in output over the business cycle, there was no long-run effect on output or the real rate, though this result really emanates from the built-in assumptions, but only on the inflation rate. Money demand shifts had a much faster impact on prices than money supply shifts and significant impact on real balances, but little influence on output variability or the real rate.

The impact of the IS shocks on output started within the same quarter as the shock, clearly just a result of the assumptions made, reached a peak two quarters after the shock but almost vanished after four quarters. However, such shocks had permanent effects on money growth, inflation and the nominal rate. While they increased the nominal rate, they first increased the real rate but, because of their impact on inflation, soon led to a decline in the real rate, which returned to its initial level about two years after the shock. IS shocks accounted for a substantial part of the business fluctuations.

Gali's findings, therefore, support the Keynesian conclusions that demand shocks do cause variations in output. Further, money supply variations had a longer-lasting impact on output than IS shocks. However, the major source of the variations in output was supply rather than demand shocks. Demand shocks did not produce long-run effects on output and the real rate of interest.[14] While Gali's model did not distinguish between positive and negative

[14] Note that these findings are affected by the assumptions made to segment the shocks to the economy, so that any errors in these assumptions could lead to erroneous results. For example, if aggregate demand changes do cause long-run changes in output, such impact would have been erroneously attributed to supplyfactors.

money supply shocks or test for asymmetry in their effects, it can clearly be modified to do so.

In another study, for a large number of countries, Bullard and Keating (1995) used vector autoregression to investigate the impact of inflation on output. In their procedure, they identified the permanent component of inflation as being due to permanent changes in the money growth rate while exogenous shocks to output were taken to cause only transitory shocks to inflation. Their finding was that permanent shocks to inflation did not *permanently* increase the level of output for most of the countries but did do so for certain low-inflation countries. In general, the estimated effects were positive for low-inflation countries, and low or negative for high inflation countries. Money is not neutral under these findings. This pattern of the effects of inflation on output seems to reflect the opinion of most monetary economists.

A compact form of the closed-economy new Keynesian model

The following presents a compact model of the new Keynesian model along the lines discussed in Chapter 15 (Bernanke and Woodford, 1997; Clarida *et al.* (Clarida, Gali and Gertler, CGG), 1999; Levin *et al.* 1999, 2001). It has three core equations: the IS equation, the output supply equation and the monetary policy rule.

IS equation:

$$x_t = E_t(x_{t+1}) - \psi(R_t - E\pi_{t+1}) + g_t \qquad (75)$$

where $x \;\; y \;\; y^{\mathrm{f}}$, r is real interest rate while R is the nominal rate, π is the inflation rate and g represents all sources of expenditure (e.g. government deficits) other than investment. Optimizing forward-looking consumers and firms, the Fisher equation for perfect capital markets (so that $r_t \; R_t \; E\pi_{t \;\; 1}$) and rational expectations have been incorporated in this derivation of the IS equation. +

Price adjustment process ("new Keynesian Phillips curve"):

$$\pi_t = \alpha x_t + \beta E_t \pi_{t+1} + v_t \qquad (76)$$

where x is the output gap, acting as a proxy for marginal cost, mc, which rises with an increase in output. Firms rationally anticipate future inflation and smooth price adjustments. $v_t \; \rho v_{t \;\; 1} \; \eta_t$ and η is a random variable with a zero mean and constant variance. Besides other sources of disturbances, it can encompass deviations from the linear impact of the output gap on marginal cost.[15]

The preceding price adjustment equation can be rewritten as the output supply equation:

$$x_t = (1/\alpha)\pi_t - (\beta/\alpha)E_t\pi_{t+1} - (1/\alpha)v_t \qquad (77)$$

In this format, the output gap responds to both current and future inflation.

[15] See Gali and Gertler (1999), Clarida *et al*. (1999, p. 1667, fn 15).

The central bank's monetary policy rule:

The central bank derives its optimal interest rate rule by minimizing a quadratic loss function over inflation and the output gap. This yields the central bank's real interest rate rule as:

$$r^{T}{}_t = r^{LR} + \lambda x_t + \beta(E_t \pi_{t+1} - \pi^{T}) \qquad \lambda, \beta > 0 \tag{78}$$

where r^T is the target interest rate and r^{LR} is the long-run interest rate.

The model consisting of equations (75), (77) and (78) clearly does not have short-run neutrality of monetary policy; a change in the interest rate by the central bank changes aggregate demand in the IS equation, which changes the inflation rate, which, in turn, changes the output gap. Since firms simultaneously determine their output and prices in response to changes in demand, an alternative interpretation of the sequence of effects would be: a change in the interest rate by the central bank changes aggregate demand in the IS equation, to which the response by firms changes their output and marginal costs, which leads to changes in their prices, so that the output gap and inflation change.

Long-run output and inflation:

In the long run, output is at its full-employment level which, by definition, is the long-run equilibrium level of output. Hence, by the assumptions for the long run, the output gap x^{LR} is zero, so that the long-run price adjustment equation becomes:

$$\pi^{LR}_t = \beta E_t \pi^{LR}_{t+1} + v_t \tag{79}$$

Note that with long-run output always equal to its full-employment level, which is independent of aggregate demand and its determinants, output is invariant with respect to monetary policy. Therefore, monetary policy can only affect the inflation rate. It affects the current inflation rate through the expected future rate, which depends on monetary policy and its credibility. In particular, a credible monetary policy with a target inflation rate of π^T will ensure that future inflation will be at this rate.[16]

Empirical findings on the new Keynesian model

Chapter 15 has already discussed the empirical validity of the new Keynesian model in general, and that of its three components. The following adds to the evidence discussed there and focuses on the components individually.

Empirical findings on the Taylor rule

The empirical validity of the Taylor interest rate rule depends on the primary monetary policy instrument used by the central bank. Chapter 13 had argued that it cannot a priori be taken for granted that the central bank sets the money supply or the interest. Some central banks set one and some the other one. For central banks that set the interest rate, there is by now substantial evidence that they follow some form, though often with time-variant coefficients, of the

[16] Stationarity of the inflation rate in the long run requires that current inflation and future inflation be equal on average, which would require that $\beta = 1$.

Taylor rule, even when they do not explicitly announce this as their practice (see Chapter 15, Section 15.6). In general, the likelihood of a finding of the Taylor rule is very high for central banks that attempt to stabilize aggregate demand and control inflation through interest rates. However, since the coefficients of the Taylor rule depend on central bank preferences and the economy's constraints, there will be differences in the estimated coefficients of the Taylor rule among countries and even among different monetary policy regimes (e.g. with a different head of the central bank) of any given country.

In any case, central banks do not explicitly disclose their form of the Taylor rule and its coefficients for the output gap and the deviation of inflation from its target level, so that it needs to be estimated from *ex post* data. While some form of the Taylor rule does quite well empirically, its consistency with the new Keynesian models requires a forward-looking version. Levin *et al*. (1999, 2001) present the estimated coefficients of the Taylor rule from several studies, with the coefficients, and sometimes the rule itself, differing quite significantly between the studies. They conclude that a simple version of the inflation and output-targeting rule for the US economy performs quite well for the USA, and that a robust policy rule includes responses to a short-horizon forecast of inflation, not exceeding one year, and the current output gap. It also incorporates a high degree of policy inertia.

Maria-Dolores and Vazquez (2006) compare the performance of four (contemporaneous, backward-looking, a priori forward-looking, and the forward-looking rule derived from the central bank's optimization of its loss function) types of Taylor rule. They report that the new Keynesian model does much worse with a rule derived from the central bank's optimization than the backward-looking and the simple a priori forward-looking rules. Further, a simple autoregressive model of the interest rate can sometimes do better than the Taylor rule, as Depalo (2006) reports for Japan.

Empirical findings on the new Keynesian price adjustment equation

Rudd and Whelan (2003) test the general form of the forward-looking NK output equation:

$$\pi_t = \lambda x_t + \beta E_t \pi_{t+1} + v_t \qquad (80)$$

where *x* can be specified as the output gap or the deviation of the unemployment rate from the natural one. They report that this equation performs very poorly for USA data. Empirically, current inflation is negatively, not positively, related to the future output gap; intuitively put, inflation is a negative leading indicator of future output. One reason for such a result could be that their proxy used for full-employment output is a poor one. A second reason could be that, since marginal cost is unobservable, the output gap in the NKPC was used as a proxy for real marginal cost (i.e. the ratio of marginal cost to price), but may be a poor proxy. Another proxy that has been suggested for the real marginal cost is labor's share in income (Gali and Gertler, 1999), but this also has not given much better results.

A third reason for the poor performance of the NKPC could be the high degree of persistence in inflation, which depends heavily on its own lagged values (Rudd and Whelan, 2003; Maria-Dolores and Vazquez, 2006). To capture this persistence, one suggestion is to replace the NKPC by a hybrid rule, say of the form:

$$\pi_t = \delta_1 \sum_{i=1}^{n} \pi_{t-i} + \alpha \sum_{j=0}^{\infty} \beta^j E_t x_{t+j} \qquad (81)$$

A simpler form of this equation would use, for the lagged terms, only last period's inflation rate. However, for the NKPC, the justification for the backward-looking, rather than the forward-looking, inflation term on the right-hand side is problematical: for one thing, interpreting (81) as a form of (80) would be a denial of rational expectations. Further, since persistence plays a big part in inflation, the estimate α may not prove to be significant, so that (81) would reduce to just a form of static expectations. However, even if α proves to be significant, the estimated contribution of the forward output gaps may prove to be relatively small. Moreover, the estimated coefficients of this equation may shift with policy, so that the Lucas critique applies to them.

Further, as mentioned in Chapter 15, Mankiw (2001) provides three inconsistencies between this sticky price adjustment equation and the stylized facts about the relationship between inflation and unemployment. Among these inconsistencies are that this process does not generate persistence in inflation but does generate implausible dynamic adjustments in inflation and unemployment in response to monetary shocks. This study reports that the relationship that best fits the facts is the backward-looking one:

$$\pi_t = \beta\pi_{t-1} + \alpha(u_t - u^{\mathrm{n}}) + v_t \tag{82}$$

which is closer to the traditional Phillips curve, or to one with static or adaptive expectations. To provide a theoretical justification consistent with the current intertemporal optimizing approach in new Keynesian economics, Mankiw and Reis (2002, 2006a,b) replace the sticky price process by a sticky information one. The latter implies a backward-looking inflationary process which generates persistence in inflation, so that there is a significant difference between the sticky price adjustment and the sticky information equations of the Phillips curve.

Ball's Keynesian small open-economy model with a Taylor rule

Ball (1999, 2000) presents the following compact Keynesian model, with all variables in logs, for a small open economy:

IS equation:

$$y_t = -\beta r_{t-1} + \delta e_{t-1} + \lambda y_{t-1} + \varepsilon_t \tag{83}$$

Phillips curve:

$$\pi_t = \pi_{t-1} + \alpha y_{t-1} + \gamma(e_{t-1} - e_{t-2}) + \eta_t \tag{84}$$

Exchange rate determination:

$$e_t = \theta r_t + v_t \tag{85}$$

In these equations, *y* is the log of output, *r* is the real interest rate, *e* is the log of the real exchange rate[17] (a higher value of *e* means appreciation of the domestic currency), π is the current inflation rate and ε, η and v are white noise terms. All parameters are taken to

[17] For small open economies, the literature indicates that the central bank's policy function should include the

exchange rate in addition to inflation and output among its variables (see, for example, Ball, 1999, 2000), though rules excluding the exchange rate also seem to do quite well.

be positive. (83) is an open-economy IS equation, with commodity demand depending on the (lagged values of the) real interest rate and the real exchange rate. (84) is an open-economy backward-looking price/inflation adjustment relationship, with the change in the inflation rate a function of lagged output and the lagged change in the exchange rate. The change in the exchange rate affects inflation through imports since depreciation increases import prices, which increases the domestic price level. (85) has a negative relationship between the interest rate and the exchange rate; an increase in the domestic interest rate increases capital inflows into the domestic country, which causes an appreciation of the domestic currency (i.e. the exchange rate rises). This model clearly embodies the short-run non-neutrality of aggregate demand and therefore of monetary policy, and is consistent with the new Keynesian sticky price hypothesis.

The preceding model is missing an equation for the money market. Such an equation can be the LM equation under the assumption of an exogenous money supply or an interest rate policy rule, such as the Taylor rule. Ball assumed that the central bank sets the real interest rate as:

$$r_t = a_0 r_{t-1} + a_\pi [\mathrm{E}(\pi_{t+4} | I_t) - \pi_t{}^*] + a_y y\mathrm{gap}_t + \mu_t \tag{86}$$

where $\pi_t{}^* \;\; \pi_t \;\; \gamma \;\; e_{t\;\,1}$; that is, the central bank makes the desired long-run inflation rate invariant to changes in the exchange rate by filtering out the impact of lagged exchange rate changes on the current inflation rate.[18] *y* gap is the output gap, defined as the deviation of output from its full-employment level.

Results of other testing procedures

The findings in numerous empirical studies have ranged back and forth against the neutrality assumption. We do not intend to review many more studies but do consider the findings of a different type of study to be worth mentioning. As against the use of compact (small) reduced-form models reported above from the works of Barro, Mishkin, Gali, and Bullard and Keating, Mosser (1992) based her findings on four large structural macroeconometric models, well established in the late 1980s and 1990s.[19] These were used to generate the elasticities of real output, real interest rates and various other real variables with respect to monetary variables such as M1 and non-borrowed reserves held by banks. The estimated elasticities were significant not only for the first four quarters but for periods longer than 12 quarters for many of the real variables. For real output, the elasticities were positive and continued to increase up to 12 quarters, and were significant (either positive or negative) even at 40 quarters. Hence, not only were the monetary variables not neutral, the estimated lags were very long.

Summing up the empirical evidence on monetary neutrality and rational expectations

The results reported from Mishkin (1982), Gali (1992), Mosser (1992) and Ball (1999, 2000), as against Barro's (1977), supported the Keynesian theory and rejected the modern classical

[18] To illustrate, for China, Wang and Handa (2007) estimate (83) to (85) by a simultaneous equations technique and estimate the Taylor-rule equation (86) using cointegration and error-correction modeling. They find support for both this rule and the above model for China, a developing economy.

[19] These were: the Bureau of Economic Analysis model, the Data Resources Inc. model, the Federal Reserve Board/MPS model and the Wharton Econometric Forecasting Associates model.

one on the critical difference between them on the neutrality of anticipated money supply changes and the continual clearance of all markets. This implies rejection of the Friedman–Lucas supply rule, which incorporates the neutrality of anticipated money supply changes. In general, in spite of its optimizing microeconomic foundations and their intellectual appeal, the modern classical macroeconomic models have not been an unqualified empirical success.

Another aspect of neutrality modeling worth noting is that the supply functions used embody either neutrality (for classical models) or non-neutrality (for Keynesian models) of systematic monetary policy. However, empirical evidence suggests that systematic monetary policy can be neutral sometimes and non-neutral at other times, and that the "degree of non-neutrality" can be variant in intermediate cases, though it is a priori difficult to separate these cases (Lucas, 1994, 1996). These results need not come as a big surprise; the various theories presented in Chapters 14 and 15 indicate that there can be very many different causes of non-neutrality in the economy: staggered wage contracts and the lagged adjustment of nominal wages to prices; disequilibrium factors in the neoclassical model; deficient demand in the Keynesian model; relative price errors in a Lucas-type model; sticky prices in certain markets, sticky information and sticky pensions and other predetermined sources of incomes, etc. Consequently, if our estimating equation allows only the black/white scenario of neutrality versus non-neutrality, when part of the data is from a neutral sample and part is from a non-neutral one, we are likely to get a mixed bag of empirical results, varying with the relative weights of the two types of data in our sample.

Overall, while there is a somewhat mixed bag of empirical studies favoring one or the other of these hypotheses, the empirical evidence seems more often to favor the short-run non-neutrality of money rather than its neutrality. The evidence also seems to favor the finding that there is a difference between the effects of rationally anticipated *money supply* changes and those of unanticipated *price* changes, but both can have short-run effects on real output.

By comparison, the empirical evidence on the rationality of expectations does not, in general, reject it. In any case, its assertion that economic agents take account of all available information rather than merely that on the past, seems to be incontrovertible. However, at the level of implementation, whether this hypothesis justifies the interpretation of the modern classical approach that the expected values for the next few quarters or even years are the long-run equilibrium ones is highly doubtful. The more realistic interpretation clearly seems to be the Keynesian one: the rationally expected values are related to the actual future values, which depend on the actual performance of the economy, which is not necessarily the full-employment one, and the stage of the business cycle.

Getting away from dogma

The LSW model makes the implicit assumption, derived from perfectly competitive and efficient markets, that changes in output occur only in response to changes in prices. As some of the models argue, competitive markets do not necessarily have instantaneous restoration of prices to their equilibrium levels after a shock. If markets are slow to adjust but economic agents react faster to changes in demand or supply shocks, economic agents may change their output, employment, consumption and investment without a prior (or "full") change in prices and wages. Given the sluggish adjustment of prices and wages by markets, appropriate models for the economy must also consider the possibility of firms' responses to actual and expected changes in demand and workers' responses to the expected and actual changes in employment. In such a context, output and employment may respond to policy measures without the impact of these policies first occurring through a price change. Chapter 14 supports this proposition by a quote from Robert Lucas, Jr, that is worth repeating here:

> Sometimes, as in the U.S. Great Depression, reductions in money growth seem to have large effects on production and employment. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so.
>
> (Lucas, 1994, p. 153).
>
> Anticipated and unanticipated changes in money growth have very different effects. [However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that] *only small fractions* of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, *they donotseemtobetransmitted through price increases*, as in Lucas (1972).
>
> (Lucas, 1996, p. 679, italics added).

These quotes, and the inconsistency of the implications of the Friedman–Lucas supply equation with the stylized facts set out at the beginning of Chapter 14, are convincing evidence that this supply equation is not valid for most or all modern economies.

The output equation revisited

Our preceding arguments and empirical assessment can be captured by using the identity:

$$y \equiv y^{\mathrm{f}} + (y* - y^{\mathrm{f}}) + (y - y*) \tag{87}$$

where y is the actual output, y^{f} is the long-run output in the absence of errors in expectations, and y^* is the short-run equilibrium unemployment rate in the presence of errors in expectations. The Friedman and Lucas supply analyses (see Chapter 14) imply that:

$$y* - y^{\mathrm{f}} = f(P - P^{\mathrm{e}}) \tag{88}$$

Therefore,

$$y = y^{\mathrm{f}} + f(P - P^{\mathrm{e}}) + [(y - y*)|OY/OP] \tag{89}$$

where $[(y\ y^*)\ OY/OP]$ is meant to indicate the deviation of actual output from the short-run value, resulting from errors in price expectations in the labor and commodity markets. Hence there are two reasons for deviations from full-employment output. One of these occurs through price changes. For these deviations, the Friedman and Lucas supply analyses imply that, for perfect markets and rational agents, it is not the price change itself that is relevant but the deviation of the price level from its expected value. The second reason for deviations arises from the changes in aggregate demand or supply that do not proceed through price level changes. These are captured through the term listed as $[(y\ y^*)\ OY/OP]$, which has numerous potential causes and may occur in some stages of the economy but not in others. If none of them operate in a particular context, this term would be zero, so that anticipated changes in demand induced by anticipated monetary and fiscal policies would not have any impact on output. But, as the Keynesians argue, the term need not be zero in all potential cases. Hence, expansionary monetary and fiscal policies may change the price level but their impact on output is unlikely to be fully reflected through price changes. In Lucas's assessment, only a fraction is so reflected, so that the larger part of the impact of money supply changes on output seems to occur through the term $[(y\_y^*)|OY/OP]$, whose determinants need to be specified. The Keynesian paradigm indicates several of them, but there may still be others.

While Keynesian models allow the relationship between output and inflation in period t to depend on expected inflation in future periods ($t + 1$, and so on) they do not incorporate the determinants and effects of errors in expectations $(P_t - P^e{}_t)$ or $(\pi_t - \pi^e{}_t)$ during the current period. Classical models do so but suffer from their failure to incorporate the determinants and effects of $[(y\ y^*)\ \Phi Y/OP]$ in their models. Addressing these deficiencies in a single model may help in addressing Lucas's criticism, quoted above, of existing macroeconomic models: there is no generally accepted model in which reductions in money growth sometimes "seem to have large effects on production and employment" while, at other times, "large reductions in money growth seem to have been neutral, or nearly so" (Lucas, 1994, pp. 153–4).

The Phillips curve revisited

For explaining the observed levels of unemployment, the preceding arguments imply that the appropriate form of the Phillips curve should be derived from the identity:

$$u \equiv u^n + (u* - u^n) + (u - u*) \tag{90}$$

where u^n is the natural rate of unemployment and u* is the short-run equilibrium unemployment rate in the presence of errors in price expectations. From the Friedman and Lucas analyses, we have,

$$(u* - u^n) = g(\pi - \pi^e) \tag{91}$$

Therefore,

$$u = u^n + g(\pi - \pi^e) + [(u - u*)|OY/OP] \tag{92}$$

The term $[(u\ u^*)\ OY/OP)]$ means that $(u\ u^*)$ is conditional on changes in real aggregate demand. If the price level changes do fully reflect the change in nominal aggregate demand, real aggregate demand will not change, so that $[(u \quad u^*)\ OY/OP]$ will be zero. But if real aggregate demand does change, the unemployment rate will change. To illustrate, the impact of an expansionary monetary policy on aggregate demand, which does not proceed through a change in the price level, will change the actual unemployment rate. The new Keynesian models provide reasons why the third term on the right-hand side of (90) need not be zero.

The implication of the second term on the right-hand side of (90) for the simple Phillips curve (stated as $u = f(\pi)$) is that the Phillips curve for a period t will shift if the inflation rate expected for period t changes. While the new Keynesian models do incorporate the effect of future inflation on the relationship between current inflation and current output, they do not incorporate the effect of errors between the inflation rate in period t and the inflation rate *expected* for period *t* itself. While the classical economics models of Lucas and Sargent and Wallace incorporate this element, they do not incorporate the third term on the right-hand side of (90). We conclude that the appropriate Phillips curve needs to incorporate elements of both the modern classical analysis and the Keynesian analysis, with plenty of possibility of revisions, corrections and new theories to fill in the gaps.

Hysteresis in long-run output and employment functions

The new Keynesian and the modern classical approaches agree that long-run employment and output are independent of inflation and aggregate demand, which also makes them independent of the short-run performance of the economy. This implies the absence of hysteresis, which is broadly defined as the impact of the short run on the long-run performance of the economy. While this absence is analytically taken for granted and is hard to establish empirically, there is some evidence that long-run, or at least long-term, employment depends on the duration and extent of booms and recessions in the economy, and hence on short-run aggregate demand factors.[20]

Conclusions

There are considerable disputes on the underlying macroeconomic model of the economy. The classical propositions of the Friedman–Lucas supply rule propose a relationship in which output is invariant to anticipated inflation and anticipated money supply, while it may not be invariant to the unanticipated values of these variables. There is considerable evidence by now that anticipated monetary policy is not neutral, at least in the short run. This rejection of the central doctrines of the modern classical school opened the way in the 1990s for the resurgence of Keynesian approaches, culminating in the new Keynesian model. This model follows the agenda of the modern classical school that macroeconomic theory has to be based on microeconomic foundations with rational expectations. The core parts of this model are its forward-looking price adjustment equation and the Taylor rule for interest rate setting. Unfortunately, both of these have been strongly rejected in favor of their backward-looking versions.

This chapter has also suggested that the equations for output and unemployment need to encompass several sources of deviations from full employment. One of these sources is errors in expectations, incorporated in the modern classical models, but there are also many other such sources, some of which are incorporated in the Keynesian and new Keynesian models. Further, in addition to the fact that the impact of monetary policy is not neutral, a considerable part of its impact on output and unemployment does not go through price level changes.

Assessment on the choice of the variable (money supply or interest rate) that should be exogenously set by the central bank:

> We didn’t abandon the monetary aggregates, they abandoned us.
>
> (Gerald Bouey, Governor of the Bank of Canada in the late 1980s).

[20] For instance, Mankiw (2001) suggests a relationship in which the long-run unemployment is a function of the actual short-term unemployment rate.

The profession's assessment on the neutrality of money and discretionary monetary policy:

In their summing up of the effect of monetary changes on output, Milton Friedman and Anna Schwartz (1963) wrote the following:

On the non-neutrality of money:

> Three counterparts of such crucial experiments [of physical science] stand out in the monetary record since the establishment of the Federal Reserve System. On three occasions, the System deliberately took policy steps of major magnitude which cannot be regarded as necessary or inevitable economic consequences of contemporary changes in money income and prices. Like the crucial experiments of the physical scientist, the results are so consistent and sharp as to leave little doubt about their interpretation. The dates are January–June 1920, October 1931, and July 1936–January 1937. These were three occasions … when the Reserve System engaged in acts of commission that were sharply restrictive … each was followed by sharp contractions in industrial production … declines within a twelve-month period of 30 per cent (1920), 24 per cent (1931), and 34 percent per cent (1937), respectively.
>
> (Friedman and Schwartz, 1963, pp. 688–9).

The magnitude of the real effects in these examples is remarkable. Lest it be thought that only monetary contractions have real effects, Friedman and Schwartz (1963, p. 690) also cited three very significant episodes where monetary expansions by the Federal Reserve System caused large increases in industrial production.

Laurence Ball and N. Gregory Mankiw (1994), two of the foremost new Keynesians, write:

On the non-neutrality of money:

> We believe that monetary policy affects real activity. The main reason for our belief is the evidence of history, especially the numerous episodes in which monetary contractions appear to cause recessions. … monetary contractions are a major source of U.S. business cycle.
>
> (Ball and Mankiw, 1994, pp. 128–9).

On the nineteenth-century classical and Friedman's views on the neutrality of money:

> The [pre-Keynes] classical economists never suggested that money was neutral in the short run [but did offer] the key insight … that money is neutral in the long run.
>
> (Ball and Mankiw, 1994, p. 132).

On the sources of the non-neutrality of money:

> We believe that price stickiness is the best explanation for monetary non-neutrality… many prices change infrequently… (though) many (other) prices in the economy are quite flexible. … Other economists, however, accept monetary non-neutrality but resist

> the assumption of sticky prices. They have been led to develop models of non-neutrality with flexible prices.
>
> (Ball and Mankiw, 1994, pp. 131, 134–5).

On the influence of monetary policy and the role of the central banks, Ball and Mankiw assert that:

> The Fed is a powerful force for controlling the economy…. Policymakers and the press believe that monetary policy can speed up or slow down real economic activity.
>
> (Ball and Mankiw, 1994, 132–3).

On the new Keynesian model

The new Keynesian model is the latest of the macroeconomic models to appear on the scene. It differs from the earlier Keynesian approaches by explicitly deriving its components from microeconomic, intertemporal foundations and rational expectations. In this, it follows the pattern of the modern classical models but differs from the latter by its addition of market imperfections and price stickiness. As Mankiw (2001) points out:

> [Since the time of David Hume, it has been well known that] a monetary injection first increases output and inflation, and later increases the price level (p. C46).
>
> The so-called "new Keynesian Phillips curve" is appealing from a theoretical standpoint, but it is ultimately a failure. It is not at all consistent with the standard dynamic effects of monetary policy, according to which monetary shocks have a delayed and gradual effect on inflation. We can explain these facts with traditional backward-looking (original Phillips curve) models of inflation–unemployment dynamics, but these models lack any foundation in the microeconomic theories of price adjustment.
>
> (Mankiw, 2001, p. C52).

Further, the new Keynesian derivation of the monetary policy rule in the form of a forward-looking Taylor rule does not fare any better when compared with a backward-looking Taylor rule.

How do the above views of Keynesians compare with those of Lucas (1994), who is associated with the modern classical school and has been a major contributor to it? Some of his views are:

On the variant neutrality and non-neutrality of money:

> Sometimes, as in the U.S. Great Depression, reductions in money growth seem to have large effects on production and employment. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so. Observations like these seem to imply that a theoretical framework such as the Keynes–Hicks–Modigliani IS/LM model, in which a single multiplier is applied to all money movements regardless of their source or predictability, is inadequate for practical purposes.
>
> (Lucas, 1994, p. 153).

On the lack of adequate knowledge and absence of theories of the variant non-neutrality of money:

> Little can be said to be firmly established about the importance and nature of the real effects of monetary instability, at least for the U.S. in the postwar period. Though it is

widely agreed that we need economic theories that capture the non-neutral effects of

> money in an accurate and operational way, none of the many available candidates is without serious difficulties. (p. 153).
>
> Macroeconomic models with realistic kinds of monetary non-neutralities do not yet exist (1994, pp. 153–54). … anticipated and unanticipated changes in money growth have very different effects (1996, p. 679).
>
> [However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that] only small fractions of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, they do not seem to be transmitted through price increases, as in Lucas (1972).
>
> (Lucas, 1996, p. 679).

In the "Nobel Lecture" (1996), given on his receipt of the Nobel Prize in economics, Robert Lucas added that:

> In summary, the prediction that prices respond proportionately to changes in the long run, deduced by Hume in 1752 (and by many other theorists, by many different routes, since), has received ample – I would say decisive – confirmation, in data from many times and places. The observation that money changes induce output changes in the same direction receives confirmation in some data sets but is hard to see in others. Large-scale reductions in money growth can be associated with large-scale depressions or, if carried out in the form of a credible reform, with no depression at all.
>
> (Lucas, 1996, p. 668).

A central banker's opinions on the neutrality of money, lags and the art of monetary policy

What do central bankers, who are the real-world practitioners of monetary policy, in fact believe and do? In 1997, the central banks of both Canada and the USA had their declared objective as that of aggressively promoting price stability. Both used interest rates, rather than monetary aggregates, as their target and instrument variables. On the dynamics of their policies, a speech by the Governor of the Bank of Canada on October 7, 1997, expressed the stance of the Bank's policy as:

> Too much monetary stimulus can lead to an exhilarating temporary burst of economic activity. But it will almost certainly also lead to inflation-related distortions that undermine both the expansion and the economy's efficiency over the longer term. The end-result, as we know only too well from past experience, is high interest rates, punishing debt loads, recession, and higher unemployment.
>
> A further complication is that it takes between a year to a year and a half for the economy to fully respond to changes in the degree of monetary stimulus … you want to look way ahead to see what is coming, and you want to take action early … That is why monetary policy must focus on future, rather than the present, and why the Bank must act in a forward-looking pre-emptive manner.
>
> (Gordon Thiessen, Governor of the Bank of Canada, 1997)[21].

[21] "Challenges ahead for monetary policy." *Remarks to the Vancouver Board of Trade*, 7 October1997.

Collectively, these assessments of economists from the Keynesian and classical traditions and of a central banker show a high degree of agreement that, in the short run, money can be non-neutral, and more likely to be non-neutral than neutral. However, ignoring the possibility of hysteresis, there is also broad agreement and substantial evidence that long-run output growth is independent of money supply growth.[22] While these conclusions indicate much less divergence on the nature of the economy than the formal models of the different schools convey, there is little agreement on the sources and extent of the potential short-run non-neutrality. There can be several possible sources of non-neutrality, as discussed in the first part of these conclusions, and the reasons for and the extent of non-neutrality can differ at different times and in different countries.

Summary of critical conclusions

- ❖ For many economies, the central bank's control of the interest rate as the variable of monetary policy provides a better tool for management of aggregate demand in the economy.
- ❖ The LSW model, incorporating the Friedman–Lucas supply function and rational expectations, is often used as the compact form of the modern classical model for short-run macroeconomics. Its policy recommendation is that the central bank should not use changes in the money supply to attempt changes in output and unemployment in the economy.
- ❖ While a negative relationship between the rate of unemployment and the rate of inflation seemingly occurs in the LSW model, the coefficients of such a relationship are not invariant to shifts in monetary policy. The Lucas critique would apply to such a relationship.
- ❖ The LSW model modified by the Keynesian supply function, as well as the new Keynesian models, allows systematic monetary policy to change output and unemployment in the economy.

Review and discussion questions

1. What evidence would you need to establish whether or not money supply changes have been the main cause of changes in nominal income? What procedure can you use to determine the direction of causality between the changes in money and in income?
2. Specify the hypotheses on the natural rate of unemployment and the rational expectations. Discuss the logical and historical relationship between them.

   Discuss, for each concept, whether disequilibrium in the economy is consistent with it or not. If it is, discuss the role and usefulness of monetary policy in this state.
3. Discuss the evolution of the 1970s Monetarists' claim that "only monetary policy is effective" to the doctrines of the modern classical school that "no foreseen monetary policy is effective" and "no systematic monetary policy is effective." Would Friedman have subscribed to any of these propositions?

[22] This does not imply that output growth is independent of innovations in the financial sector. For this analysis, see Chapter 24.

4. Inflation and unemployment are two crucial economic items of interest to the public. Is it possible to explain one without the other? Present at least one theory that explains each independently of the other and one theory that establishes their interdependence. Is there a genuine difference between these theories or do they merely represent a distinction between the impact and long-term effects of an exogenous change to their determinants?
5. Present a model with rational expectations and the Friedman–Lucas supply function. If policy makers and the public have the same information, can stabilization policies in a stochastic context change aggregate demand and output (i) in the short run, (ii) in the long run?
6. Present a model with rational expectations and the new Keynesian supply function. If policy makers and the public have the same information, can stabilization policies in a stochastic context affect aggregate demand and output (i) in the short run, (ii) in the long run?
7. Why do models with rational expectations have difficulty in explaining the persistence of output from its trend and unemployment from the natural rate? What are some of the reasons given for this persistence? If this persistence were incorporated in them, what would be their implications for the effectiveness of monetary policy: could activist monetary policy stabilize output and the unemployment rate? Discuss in the context of a specific model embodying such persistence.
8. Outline the development and current theoretical status of the tradeoff between the unemployment rate and the rate of inflation.
9. "Between 1930 and 1990, macroeconomic theory went through a full circle. The classical view of the 1920s was discarded in the 1930s and 1940s by Keynesian theory, but the latter, in turn, was gradually eroded to the point that the dominant theory by the 1980s had again become a form of the classical one." Discuss.
10. "In the past forty years, macroeconomic theory has come full circle. The Keynesian views of the economy were discarded in the mid-1970s by the resurgent classical theory, but the latter, in turn, has been eroded to the point that the dominant theory is again a form of the Keynesian one." Discuss.
11. Modern classical macroeconomics argues that anticipated monetary policy does not have real effects. The new classical macroeconomics argues that anticipated fiscal policy also does not have real effects. Adapt the Lucas–Sargent–Wallace model to explicitly incorporate both of these propositions. From this model, what would you conclude for the effects of (i) an anticipated bond-financed deficit, (ii) an anticipated money-financed deficit? Specify your procedure and estimating equations for testing the validity of your conclusions.
12. Consider an economy with the following structure: Aggregate demand:

$$y_t = M_t - P_t + \mu_t \qquad \text{(A quantity theory type equation)}$$

Aggregate supply:

$$y_t = y^f_t + \gamma\,(P_t - P^e_t) + \eta_t$$

where the symbols have the usual meanings and are in logs. μ and η are random errors.

Expectations are formed in one of the following alternative ways:

(a) Rational expectations:

$$P^{\mathrm{e}}{}_t = E_{t-}\ P_t$$

(b) Adaptive expectations:

$$P^{\mathrm{e}}_t - P^{\mathrm{e}}_{t-} = (1-\lambda)(P_{t-} - P^{\mathrm{e}}_{t-})$$

(i) Given rational expectations, solve for P_t and y_t (in terms of the money supply and random errors, etc.).
(ii) Given adaptive expectations, solve for P_t and y_t (in terms of the money supply, etc.).
(iii) For the two expectations hypotheses, derive the time patterns of the response of the price level to a unit change in the money supply.
(iv) For the two expectations hypotheses, derive the time pattern of the response of real output to a unit change in the money supply.
(v) Discuss the differences implied by the two hypotheses for the impact of monetary policy on real output and prices.

13. Consider the following model:

Aggregate supply:

$$y_t = \gamma(P_t - E_{t-1}P_t) + \gamma(P_t - E_{t-2}P_t)$$

Aggregate demand:

$$y_t = M_t - P_t + \mu_t$$

$$\mu_t = \mu_{t-1} + \eta_t$$

where *y*, *P* and *M* have the usual meanings and are in logs. η is a serially uncorrelated error with mean zero and variance σ^2.

(i) How would you justify the above aggregate supply function and how does it differ from the Lucas one?
(ii) Suppose that the central bank can observe μ_{t-1} but not μ_t when it sets the money supply. Is there then a role for systematic monetary policy?
(iii) Given (ii), suppose that the central bank wants to set the money supply to minimize the variance- $\mathrm{E}_{t\ 1}(y_t\ y^{\mathrm{f}}\ )^2$ of output about its full-employment level? What monetary policy would it follow?
(iv) If the policy in (iii) has always been followed, is there any way of using econometric evidence to differentiate between the pattern shown by this economy and one in which the economy had a Lucas supply function?

14. Suppose the economy is described by: Aggregate supply:

$$y_t = y^{\mathrm{f}}$$

Aggregate demand and fiscal policy:

$$y_t = \alpha_0 + \alpha_1(M_t - P_t) + \alpha_2 E_{t-1}(P_{t+1} - P_t) + \alpha_3 z_t + \mu_t \qquad \alpha_1, \alpha_2, \alpha_3 > 0$$

$$z_t = \gamma_0 + \gamma_1 z_{t-1} + \eta_t$$

Money supply:

$$M_t = M_0 + v_t$$

where all variables are in logs. μ, η and v are random disturbance terms. y, M and P have the usual meanings, and z is a real fiscal variable.

Find the equilibrium solutions for y_t and P_t. How do systematic or anticipated changes in the nominal money supply M_0 affect y_t and P_t? How would unanticipated changes in the nominal money supply M_t affect y_t and P_t? How would anticipated and unanticipated changes in the fiscal variable affect y_t and P_t?

15. Suppose the output supply in the model of question 14 is not the full-employment level but is determined by demand. Re-do the answers to question 14.
16. Suppose the supply function in the model of this question 14 is changed to:

$$y_t = y^{\mathrm{f}} + \gamma\,(P_t - E_{t-1}P_t)\ \gamma > 0$$

Re-do the answers to question 14.

17. Suppose the price level in question 14 is fixed at $\underline{P}$ (making the model a fixed price one, compared with the preceding flexible price model). Re-do the answers to this question.
18. Specify the three types (backward-looking, contemporaneous and forward-looking) of the Taylor rule on interest rates, and discuss their validity. Why does the forward-looking form implied by the new Keynesian model seem to do badly relative to the others, and especially relative to the backward-looking one?
19. Write the general dynamic form of the sticky-price new Keynesian Phillips curve as:

$$\pi_t = \beta E_t\, \pi_{t+1} + \alpha(u_t - u^{\mathrm{n}}) + v_t$$

where π is the inflation rate, u is the unemployment rate, u^{n} is the natural rate, v_t $\rho v_{t\ \ 1}$ η_t and η is a random variable with a zero mean and constant variance. Mankiw (2001) argues that this relationship is not valid and that the general form of the valid relationship is backward-looking in inflation. Its corresponding form would then be:

$$\pi_t = \beta\pi_{t-1} + \alpha(u_t - u^{\mathrm{n}}) + v_t$$

Is this form consistent with the original Phillips curve? Provide the theoretical justification for it offered by the sticky information model, and critically evaluate this contribution.

20. Given the generally poor performance of the forward-looking forms of the Taylor rule and the Phillips curve implied by the NK model, especially relative to their backward-looking simpler versions, discuss whether the new Keynesian reformulation of

Keynesianism has proved to be a failure. If this is so, are there any elements of the NK ideas that should be preserved in further research?

References

Ball, L. "Policy rules for open economies." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: Chicago University Press, 1999.

Ball, L. "Policy rules and external shocks." *NBER Working Paper* no. 7910, 2000.

Ball, L., and Mankiw, N.G. "A sticky-price manifesto." *Carnegie-Rochester Series on Public Policy*, 41, 1994, pp. 127–51.

Barro, R.J. "Unanticipated money growth and unemployment in the United States." *American Economic Review*, 67, 1977, pp. 101–16.

Bernanke, B.S., and Woodford, M. "Inflation forecasts and monetary policy." *Journal of Money, Credit and Banking*, 29, 1997, pp. 653–84.

Bullard, J., and Keating, J.W. "The long-run relationship between inflation and output in postwar economies." *Journal of Monetary Economics*, 36, 1995, pp. 477–96.

Chu, J., and Ratti, R.A. "Effects of unanticipated monetary policy on aggregate Japanese output: the role of positive and negative shocks." *Canadian Journal of Economics*, 30, 1997, pp. 722–41.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a new Keynesian perspective."

Journal of Economic Literature, 37, 1999, pp. 1661–707.

Cover, J.P. "Asymmetric effects of positive and negative money shocks." *Quarterly Journal of Economics*, 107, 1992, pp. 1261–82.

Depalo, D. "Japan: the case for a Taylor rule? A simple approach." *Pacific Economic Review*, 11, 2006, pp. 527–46.

Friedman, M., and Schwartz, A.J. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press, 1963.

Frydman, R., and Rappoport, P. "Is the distinction between anticipated and unanticipated growth relevant in explaining aggregate output?" *American Economic Review*, 77, 1987, pp. 693–703.

Gali, J. "How well does the IS–LM model fit post-war U.S. data?" *Quarterly Journal of Economics*, 107, 1992, pp. 709–38.

Gali, J. and Gertler, M. "Inflation dynamics: a structural econometric analysis." *Journal of Monetary Economics*, 44, 1999, pp. 195–222.

Levin, A.T., Wieland, T.V. and Williams, J.C. "Robustness of simple monetary policy rules under model uncertainty." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago, 1999, pp. 263–99.

Levin, A.T., Wieland, T.V. and Williams, J.C. "The performance of forecast-based monetary policy rules under model uncertainty." *Federal Reserve System Working Paper* no. 2001-39, 2001.

Lucas, R.E., Jr. "Expectations and the neutrality of money." *Journal of Economic Theory*, 4, 1972, pp. 103–24.

Lucas, R.E., Jr. “Some international evidence on output–inflation tradeoffs.” *American Economic Review*, 63, 1973, pp. 326–34.

Lucas, R.E., Jr. “Comments on Ball and Mankiw.” *Carnegie-Rochester Series on Public Policy*, 41, 1994, pp. 153–5.

Lucas, R.E., Jr. “Nobel lecture: monetary neutrality.” *Journal of Political Economy*, 104, 1996, pp. 661–82.

Mankiw, N.G. “The inexorable and mysterious tradeoff between inflation and unemployment.”

Economic Journal, 111, 2001, pp. C45–C61.

Mankiw, N.G., and Reis, R. “Sticky information versus sticky prices: a proposal to replace the new Keynesian Phillips curve.” *Quarterly Journal of Economics*, 117, 2002, pp. 1295–328.

Mankiw, N.G., and Reis, R. “ Pervasive stickiness.” *American Economic Review*, 96, 2006a, pp. 164–9.

Mankiw, N.G., and Reis, R. “ Sticky information in general equilibrium.” *NBER Working Paper* no.

12605, 2006b.

Maria-Dolores, R., and Vazquez, J. “How does the new Keynesian monetary model fit in the

U.S. and the Eurozene? An indirect inference approach.” *Topics in Macroeconomics*, 6, 2006, article 9, pp. 1–49.

Mishkin, F.S. "Does anticipated aggregate demand policy matter? Some further econometric results."

American Economic Review, 72, 1982, pp. 788–802.

Mosser, P. "Changes in monetary policy effectiveness: evidence from large macroeconomic models."

Federal Reserve Board of New York Quarterly Review, 17, 1992, pp. 36–51.

Rudd, J., and Whelan, K. "Can rational expectations sticky price models explain inflation dynamics?" At www.federalreserve.gov/pubs/feds/2003/200346, 2003.

Sargent, T.J., and Wallace, N. "Rational expectations and the theory of economic policy." *Journal of Monetary Economics*, 2, 1976, pp. 169–83.

Wang, S., and Handa, J. "Monetary policy rules under a fixed exchange rate regime: empirical evidence from China." *Applied Financial Economics*, 17, 2007, pp. 941–50.

18 Walras's law and the interactionamong markets

This chapter focuses on Walras's law, which is fundamental to macroeconomic analysis. It underlies the IS–LM model, which has four goods – commodities, money, bonds and labor – but has explicit analysis of only three of them. The foundations of Walras's law and its validity in disequilibrium are rigorously examined in this chapter. This is also done for Say's law. The analyses of Walras's and Say's laws are followed by the derivation of their joint implications for the dichotomy between the real and the monetary sectors, and the neutrality of money.

The Pigou and real balance effects concern the impact of changes, brought about by changes in the price level, in the value of financial assets and of real balances, respectively, on the demand for commodities. They played a key role in doctrinal disputes on whether an economy functioning below full employment would return to full employment. This chapter examines their nature and empirical relevance.

These discussions are followed by analyses of effective Clower and Drèze demand and supply functions versus notional ones.

Key concepts introduced in this chapter

- A law versus an equilibrium condition
- Walras's law
- Say's law
- Dichotomy between real and monetary sectors
- Pigou effect
- Real balance effect
- Notional demand and supply functions
- Clower effective demand and supply functions
- Drèze functions

There are very few propositions in economics that have been considered to be so far beyond dispute as to be labeled "laws." Among these are Walras's law and Say's law. The former represents a constraint on all goods in the economy while the latter represents a constraint on the commodity market alone. This chapter will consider these laws for the closed economy, though the arguments can be easily extended to the open economy.

A statement in economics worthy of being called a *law* must be more than a behavioral relationship or an equilibrium condition – both of which are not necessarily valid at a given time for a given economy – since, otherwise, there would not be a special reason for assigning it a distinctive term with the compelling connotation that the word "law" possesses. Hence, while there can be several ways of defining what can be designated as a law in economics, we will define it as a statement that holds without exception under any and all conditions, so that it is an identity – or, if one is willing to be more tolerant, it must be a statement that approximates this degree of applicability. There are very few statements of this nature in microeconomics, let alone in macro or monetary economics.[1] The classical paradigm asserts that Walras's law is one of this extremely select group. However, as we discuss in Section 18.8 below, there are serious and well-founded arguments against its being a law or an identity. In particular, while it holds in the general equilibrium states of the economy, its implications for the dynamic analysis of prices, wages and interest rates need not be always valid. Hence, we conclude that it is not an identity and, therefore, not a law.

Say's law asserts that the supply of commodities creates an equal demand for them. This chapter shows that it does not hold as an identity in the modern economy with financial assets, so that it is clearly not a law for the modern economy. It really should not be a part of any modern macroeconomic analysis.

Sections 18.1 to 18.3 present the derivation of Walras's law and its implications. Section 18.4 deals with Say's law and finds it inappropriate for monetary economies. Section 18.5 discusses the implications of the joint assumption of Walras's law and Say's law for the neutrality of money and the dichotomy between the real and monetary sectors. Sections 18.6 and 18.7 present the wealth and real balance effects. Both these concepts have already been presented in Chapter 3 in the context of the Walrasian general equilibrium model of the economy. Sections 18.8 to 18.12 examine the conditions for the breakdown of Walras's law and the implications of this breakdown.

Walras's law

Walras's law is the statement that for any economy, over any given period of time, the sum of the market *values* of all the goods demanded must equal the sum of the market values of all the goods supplied. For the closed economy,[2] we define "goods" in this chapter, as in earlier chapters, to refer to commodities, money, labor (or leisure) and non-monetary financial assets ("bonds").

To explain Walras's law intuitively for a pure exchange economy, first start with the constraint on the individual's demands for goods. Assume that the individual initially possesses some commodities, some money and some bonds,[3] and that their total nominal value at current market prices is his nominal wealth ψ . He will also want to supply some labor, with a nominal labor income at current wage rates being equal to Y . Assuming that he spends this amount to acquire the commodities, money and bonds that he wants to hold

[1] Another proposition that is sometimes referred to as a law is the "law of one price," which in international trade translates to absolute purchasing power parity between countries. However, it is often rejected by empirical evidence, so that it cannot be properly regarded as a law.

Another proposition often labelled as a law is that "the demand curves for commodities slope downwards." However, this is violated for many goods, such as those with snob appeal, which have upward-sloping demand curves. Therefore, even this proposition is not an identity and not really a law.

[2] The open economy has an additional good in the form of foreign exchange held by the private and public sectors.

[3] Bond holdings can be positive (making the individual a net lender) or negative (making the individual a net borrower).

or use in the current period,[4] the total value ψ of his initial holdings of goods plus his labor income Y must equal his total expenditures on commodities, money and bonds. Since the individual's total expenditures on his purchases of commodities, money and bonds must sum to $[\psi + Y]$, the demand for any one of these three goods can be derived by subtracting his expenditures for the other two goods from $[\psi + Y]$.

If we now sum over all the individuals in the economy, the aggregate initial holdings of goods (including labor) plus the current national output becomes their supply and the aggregate expenditures on them become the value of the quantities demanded. Further, the total value of all the goods demanded must equal the total value of all the goods supplied. This is Walras's law, and it is the aggregate counterpart of the individual's budget constraint. It is often simplified to the statement that *the supply of all goods in the economy must equal the demand for all goods in the economy.*

Deriving Walras's law for a five-good closed economy

We prove Walras's law for a closed economy with a government sector and for its *five* goods – commodities, money, bonds, equities, and labor.[5] The assumption basic to Walras's law is that there is no "free" disposal of goods (i.e. economic agents do not just throw them away) so that the goods that are produced or inherited from the past are either consumed, demanded for some other reason (such as for carrying to the future as saving in the form of commodities) or exchanged against money or bonds. Assume that the closed economy has three economic agents – households, firms and the government (including the central bank) – so that there are several budget constraints to consider in the analysis. The definitions of the symbols in the constraints are given after the specification of the constraints, which are as follows.

HOUSEHOLDS' BUDGET CONSTRAINT

This constraint specifies that the payments by households for all goods (commodities, money, bonds and equities) bought must equal the funds available from the initial endowments (i.e. inherited stocks) of goods, labor income and the profits received from firms as distributed profits. That is:

$$p_c c^{\mathrm{dh}} + p_c m^{\mathrm{dh}} + p_b b^{\mathrm{d}} + p_e e^{\mathrm{d}} \equiv p_c \underline{c}^{\mathrm{s}} + \underline{M}^{\mathrm{h}} + p_b \underline{b}^{\mathrm{s}} + p_e \underline{e}^{\mathrm{s}} + W n^{\mathrm{s}} + \pi^{\mathrm{dis}} \qquad (1)$$

FIRMS' BUDGET CONSTRAINTS

Firms face three constraints. The first one specifies that the funds available to the firms from the sales of their products (to households for consumption, to other firms for investment and to government) and of new equities must equal the sum of the payments to the labor employed and funds used for the firm's investments, money holdings or distributed profits. The firms' second constraint specifies that total profits can be either distributed or retained/undistributed

[4] We have assumed that the individual is a supplier of labor services and firms are the buyers of such services.

[5] We have included three financial goods, money, bonds and equities, in the following analysis in order to show that the separate treatment of equities does not destroy Walras's law. Note that in the rest of this book and

in monetary economics generally, “bonds” are defined to include all non-monetary assets and therefore would also include equities.

by the firm. The third constraint specifies that the firm's investment must be financed from its retained earnings and the issue of new equities.

$$p_c(c^{sh} + i + g) + p_e(e^s - \underline{e}^s) + \underline{M}^f \equiv Wn^d + p_c i + p_c m^{df} + \pi^{dis} \tag{2}$$

$$\pi \equiv \pi^{dis} + \pi^{undis} \tag{3}$$

$$p_c i \equiv \pi^{undis} + p_e(e^s - \underline{e}^s) \tag{4}$$

GOVERNMENT'S BUDGET CONSTRAINT

This constraint specifies that the government must finance its deficits or surpluses by the issue of new money and bonds.

$$p_c(g - t) \equiv (M^s - \underline{M}^s) + p_b(b^s - \underline{b}^s) \tag{5}$$

The definitions of the symbols in the preceding five identities are:

p_c	price of commodities
p_b	price of government bonds
p_e	price of equities
W	nominal wage rate = rental price of labor
c^{sh}	supply of commodities to households
c^{sf}	total supply of commodities by firms $\equiv (c^{sh} + i + g)$
m	real money balances
M	nominal money stock
b	quantity of bonds
e	quantity of equities
n	number of workers (labor)
$\underline{c}$	existing stocks of commodities (held by households)
$\underline{M}$	existing nominal money stock (held by households and firms)
$\underline{b}$	existing stock of bonds (issued by the government, held by households)
$\underline{e}$	existing stock of equities (issued by firms, held by households)
g	real government expenditures on commodities
t	government's tax revenues
i	real investment by firms
π	total nominal profits of firms
π^{dis}	nominal distributed profits
π^{undis}	nominal retained (undistributed) profits of firms

The superscripts d and s stand for demand and supply respectively. The superscripts h and f stand for households and firms respectively. Underlining indicates that the value of the variable is given exogenously or was inherited from the past. Note that $c^d \equiv c^{dh} + i + g$, $c^s \equiv c^{sf} + \underline{c}^s$ and $\underline{M} \equiv \underline{M}^h + \underline{M}^f$. For simplification, (1) to (5) make the usual assumption that only firms issue equities and that only the government issues bonds.

Note that (1) to (5) are all *identities*, designated by the use of the symbol ≡, and result from the assumption that nothing is just thrown away by economic agents since doing so would be irrational. They imply that:

$$p_c(c^d - c^s) + (M^d - M^s) + p_b(b^d - b^s) + p_e(e^d - e^s) + W(n^d - n^s) \equiv 0 \tag{6}$$

where the left-hand side is the sum of the nominal excess demands for commodities, money, bonds, equities and labor. (6) is one of the ways of stating Walras's law: *the sum of the* *nominal excess demands for all goods in the economy must be zero*. (6) restated in terms of excess demands is:

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_e{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{7}$$

where $E_k{}^{\mathrm{d}}$ is the excess *nominal* demand for the kth good, $k = c, m, b, e, n.$

The distinction between Walras's law and Walrasian general equilibrium models

Walrasian general equilibrium models assume that equilibrium exists in all markets and that markets are perfect, i.e. perfectly competitive and efficient, so that they always clear. Walras's law does not embody this assumption of perfect markets, or that that any or all markets are in equilibrium. Further, Walrasian general equilibrium models make statements about the equilibrium state, which is not an identity, while Walras's law *is* an identity. Therefore, the two concepts of Walras's law and Walrasian general equilibrium models are quite different. However, Walras's law is a requirement of all Walrasian general equilibrium models, as of other models of the economy, whereas Walrasian general equilibrium is not a requirement of Walras's law.

Walras's law in a macroeconomic model with four goods

We have differentiated between bonds and equities to capture the financial structure of the economy in a more realistic manner and to show that Walras's law will hold for this structure. In the general case, it will also hold for any economy no matter how its goods are categorized. Since macroeconomic theory usually treats firms' equities and government bonds as the composite good "bonds" (see Chapter 13), we will at this stage shorten (6) in line with this usage. Walras's law for this more compact *four-good* closed economy can be stated in the following two alternate ways.

(i) $$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{8}$$

where *b* (bonds) now represents all non-monetary financial assets in the economy. (8) is the statement that *the sum of the excess demands for the four goods is identically zero*.

(ii) Equation (8) can be rewritten as:

$$p_c c^{\mathrm{d}} + M^{\mathrm{d}} + p_b b^{\mathrm{d}} + Wn^{\mathrm{d}} \equiv p_c c^{\mathrm{s}} + M^{\mathrm{s}} + p_b b^{\mathrm{s}} + Wn^{\mathrm{s}} \tag{9}$$

which is the statement that *the sum of the nominal demands for all goods identically equals the sum of the nominal supplies of all goods in the economy.*

For the general case of *K* markets, (8) and (9) generalize to:

$$\sum_{k}^{K} E_{\mathrm{d}} \equiv 0 \tag{10}$$

$k=1$

$$\sum_{k=1}^{K} x^{\mathrm{d}}_{k} = \sum_{k=1}^{K} x^{\mathrm{s}}_{k} \qquad (10^{J})$$

where x_k is the quantity of the kth good and there are K goods in the economy. (10) or (10^{J}) is the general statement of Walras's law.

Adjustment costs, which slow the adjustment in the demands and/or supplies of goods so that their short-run values differ from their long-run ones, do not change the derivation or applicability of Walras's law, nor does the introduction of uncertainty and rational expectations into the model. However, the validity of Walras's law for dynamic analysis does become questionable if the labor and commodity markets do not clear on a continuous basis. Sections 18.8 to 18.10 will examine this issue.

The implication of Walras's law for a specific market

Walras's law by itself does not assert equilibrium in each market or in any one specific market, but is a statement covering all markets in the economy. (10) implies that:

$$E_K^{d} \equiv -\sum_{k=1}^{K-1} E_k^{d} \qquad (11)$$

where $E_K{}^{d}$ is excess demand in the Kth market. This condition asserts that the excess nominal demand in the Kth market equals the sum of the excess nominal demands in the other (K– 1) markets, where we can arbitrarily designate the market for any specific good as the K th market. (11) implies that:

$$\text{If } E_k{}^{d} = 0, \text{for} k = 1,\ldots, K-1, \text{ then } E_K{}^{d} = 0 \qquad (12)$$

That is, if there exists equilibrium in K 1 markets, then there would also be equilibrium in the K th market. (12) is also sometimes used as a way of stating Walras's law. In general equilibrium analysis, (11) and (12) allow the analysis to dispense with the explicit treatment of one of the markets in the economy. Note, however, that such an omitted market continues to exist and to function but its treatment is pushed into the implicit state. Further, the solution of any of the three markets for the three prices (P, p_b, W) in (9) would be identical, irrespective of which market is omitted from the explicit analysis. Furthermore, the general equilibrium values of the real variables, such as output, employment, consumption, etc., will also be identical.

Walras's law and selection among the markets for a model

Implications of Walras's law for general equilibrium analysis

As shown above, Walras's law permits the general equilibrium conditions for any three out of the four goods in the closed-economy macroeconomic model to be explicitly specified for the solution of the overall equilibrium of the economy. Therefore, the complete model can explicitly set out the equations for any of the following four groups of markets:

(I) commodities, money and labor markets;

(II) bond, money and labor markets;

(III) commodities, bond and labor markets;
(IV) commodities, money and bond markets;

Grouping I was used by Keynes, and has become the standard set used in modern macroeconomics. The neoclassical and the modern classical models also follow this pattern. As is discussed in greater detail in Chapter 19, the traditional classical economists, prior to Keynes, had generally favored grouping II, with the bond market determining the interest rate in the loanable funds theory, the market for money determining the price level by the quantity theorem, and the labor market determining employment and – through the production function – determining output. Walras's law implies that each of the above sets would provide the same general equilibrium solution of the values of the endogenous variables, even though the explicitly specified three markets differ between the sets.

Implications of Walras's law for the dynamic analysis of markets

While the different approaches may yield the same general equilibrium values of the variables, the general pattern of economic analysis identifies a "price" variable with each market. Thus, the price level is the "price" of commodities, and microeconomic analysis identifies its determination with the demand and supply of commodities. The interest rate is similarly the one-period "price" of loans/bonds, and microeconomic analysis would identify its determination by the market for loans/bonds. The wage rate is the rental price of labor, and microeconomic analysis identifies its determination with the labor market. Hence, there is no "price" left to identify as the price of money, so that there is no price variable that can be uniquely identified with the demand and supply of money. *There can, in fact, be no such unique variable in a monetary economy since money is itself the good in which the prices of other goods (commodities, bonds and labor) are measured.* Thus, 1/*P* is sometimes said to be the value of money(in commodity units) while, at other times, the interest rate is said to be its opportunity cost, and many pre-Keynesian economists designated its value (in labor units) as 1/*Pw.*

This reasoning suggests that, if a model is to capture the dynamic movements of prices, wages and the interest rate in real-world markets, it needs to take account of the empirical reality that the price of a particular good responds *in the first instance* to the excess demand only in the market for that good. This price should not respond to the excess demands for other goods, unless these demands spill over into the excess demand for the good in question. Therefore, for dynamic analysis, the appropriate assumptions would be:

$$\partial P_t/\partial t = f\,(E_{ct}{}^{\mathrm{d}})$$

$$\partial R_t/\partial t = f\,(E_{bt}{}^{\mathrm{d}})$$

$$\partial W_t/\partial t = f\,(E_{nt}{}^{\mathrm{d}})$$

where *R* is the nominal yield on bonds. These dynamic functions seem to be consistent with common intuition and economic folklore on price adjustments in markets. Several studies of the dynamic analysis of the price level and the interest rate have in recent years followed this pattern: that is, making changes in the price level a function of the commodity market disequilibrium and making interest rate changes a function of the bond market disequilibrium[6]. Hence, for dynamic analysis, the preferred overall macroeconomic model needs to specify the commodity, labor and bond markets, while excluding the market for money balances.[7]

Walras's law and the assumption of continuous full employment

If it is assumed that the labor market is *continuously* (always) in equilibrium, $n^{\mathrm{d}} = n^{\mathrm{s}} = n^{\mathrm{f}}$

where n^{f} stands for full employment, (8) becomes,

$$E_c^{\mathrm{d}} + E_m^{\mathrm{d}} + E_b^{\mathrm{d}} \equiv 0 \tag{13}$$

The underlying assumption behind (13) is that equilibrium exists on a continuous basis in the labor market – so that $E_n^{\mathrm{d}} = 0$, by assumption – but does not do so in the commodities, money and bond markets. In fact, in economies with developed financial markets, the most plausible assumption would be that the money and bonds markets adjust the fastest to clear any disequilibrium. Regarding the commodity and labor markets, the labor markets are the slowest to adjust since they are characterized by long-term explicit or implicit contracts between the firms and their employees. In fact, one of the major disputes dividing economists into the main groupings of Keynesians and modern classical economists is precisely over the issue of whether the labor markets will or will not clear over a reasonably short period, let alone continuously.

Hence, the underlying assumption of (13) that the labor market continuously clears – while the commodity, money and bond markets do not do so – is highly questionable as a basis for macroeconomic analysis. However, while this assumption is of doubtful validity, it is often made, as in the modern classical model specified in Chapter 14.

Say's law

Say's law is attributed to the writings of Jean-Baptiste Say in the first quarter of the nineteenth century[8] and is considered to be one of the underpinnings of the traditional (pre-Keynesian) classical macroeconomic model. Its usual statement is that "*supply creates its own demand*."[9] Since this statement is meant to refer exclusively to commodities rather than to the other goods in the economy and is meant to apply only in the aggregate over all commodities, it can be more precisely formulated as: *the aggregate supply of commodities creates its own aggregate demand.*

[6] See Shaller (1983). Shaller's dynamic analysis of prices and interest rates specifies changes to these in disequilibrium in terms of the commodity and bond market equations, rather than in the commodity and money market equations.

[7] This point will be taken up again in Chapter 19 on the determination of the rate of interest.

[8] Baumol (1999) claims that the appellation "Say's law" was given in the early twentieth century to ideas of which Say was an enunciator, but that they did not originate with him. Further, Say and other writers espousing these ideas did not make claims to its being a "law."

[9] Baumol (1999) attributes this mode of statement to Keynes and states that "Keynes, at best, did not get it quite right." (p. 195). He attributes the interpretation of Say's law as an identity to Oskar Lange (1942).

Say's law was implicit in many of the expositions of the traditional classical model. It can be found in the writings not only of Say but also of Adam Smith[10] in the late eighteenth century, and David Ricardo, John Stuart Mill, Alfred Marshall and others in the nineteenth century. Chapter 1 provided some discussion of Say's law[11] and should be reviewed at this stage. The law's general implication was that there could not exist either an excess demand or an excess supply of commodities in the closed economy. The core reason given for Say's law was that the whole of saving was converted into investment, so that no part of it was converted into money holdings since the former had a positive return while the latter did not.[12]

The statement that the supply of commodities creates its own demand has two components. One is the causality from supply to demand and the other is their identical amount. On the former, the argument runs as follows: the supply of commodities creates income which the recipients must spend on commodities, so that any increase in the aggregate supply of commodities creates a corresponding increase in the aggregate demand for them. This argument is fallacious in the commodities–monetary–bonds economy, since the increase in income could be partly or wholly used to increase money or bond holdings, so that the increase in the aggregate supply of commodities would induce a less than corresponding increase in their aggregate demand. Conversely, if the economic agents choose to increase their commodity demand by running down their money or bond holdings, an increase in the aggregate demand for commodities will come about without a corresponding prior increase in their supply. Hence, the causal argument behind Say's law is not valid in modern economies.

As argued earlier, economics does not use the term "law" to refer to equilibrium conditions, otherwise the equilibrium conditions for the bonds, money and labor markets would also be referred to as laws. These conditions are, in fact, never referred to as laws. Similarly, the interpretation of Say's law as an equilibrium condition does not merit the designation of a "law." Therefore, we will henceforth treat Say's law as an identity. That is, Say's law would be interpreted as: in the aggregate, the demand for commodities *always* equals their supply, with causality taken to run from the latter to the former and not from the former to the latter.

For a rudimentary (barter) economy in which the only traded goods are commodities, Walras's law simplifies to the statement that the aggregate expenditure on commodities must always equal the aggregate income from their sale. That is, for such an economy, Walras's law implies that the demand and supply of commodities are always equal, which is identical with Say's law. Therefore, for such a rudimentary economy, Say's law can be derived as an identity from the aggregate budget constraint for the economy, obtained from summing over the budget constraints of all its economic units. Note that financial assets are excluded by assumption from such an economy, so that substitution between commodities and either of the financial assets (money and bonds) is excluded.

[10] For example, Adam Smith wrote that "What is annually saved is as regularly consumed as what is annually spent, and nearly in the same time too … saving … is immediately employed as capital either by himself or some other person. … The consumption is the same but the consumers are different." (Baumol, 1999, p. 200).

[11] There are disputes about both the attribution of Say's law and how accurately it reflected the ideas of Say and other writers who expounded it. Our concern in this book is not with the historical attribution of this statement or whether or not it accurately reflects the ideas of Say or his contemporaries, but rather with examining its implications and validity.

[12] Chapter 5 on speculative demand and Chapter 6 on the buffer stock of money imply that this reasoning is not valid for monetary economies.

Conversely, Say's law is inapplicable to an economy in which both financial assets and commodities exist and some substitution can occur between them. Hence, Say's law cannot be validly applied to modern economies, all of which have money and bonds among their traded goods.[13]

Some invalid implications of Say's law

Since Say's law is inapplicable to monetary economies, its application to such economies leads to conclusions that are objectionable and inappropriate for monetary economies. The following arguments present some of these.

1 From Say's law, since the commodity sector *always* clears irrespective of the price level of commodities, we have:

$$y^{\mathrm{d}}(P_0) \equiv y^{\mathrm{s}}(P_0)$$

as well as:

$$y^{\mathrm{d}}(\lambda P_0) \equiv y^{\mathrm{s}}(\lambda P_0)$$

for any $\lambda > 0$. Hence, the price level becomes indeterminate in so far as the commodity sector is concerned: at a price P_0, there is equilibrium and zero excess demand in the commodities market, as it is at any price $\lambda \mathrm{P}_0$. Therefore, additional information is necessary to determine the price level.[14]

2 Say's law asserts that the aggregate demand for commodities always equals their aggregate supply, irrespective of the price level. The price level is thus not affected by shifts in the commodity market. For example, an increase in investment, exports, fiscal deficits, etc., causing an increase in aggregate demand, cannot increase the price level, contrary to both intuition and the implications of almost every macroeconomic model.

3 Say's law on its own asserts that the supply and demand for commodities are always identical, irrespective of the interest rate and the level of income in the economy. Hence, in IS–LM models, the IS relationship (curve) would span the whole (r, y) space, rather

[13] Some economists define a "weak" form of Say's law as: *in equilibrium, the aggregate demand for commodities equals their aggregate supply*. This confinement of Say's law to an equilibrium condition is hardly very restrictive for the commodity market and merely becomes the specification of the IS relationship in macroeconomic analysis. As only an equilibrium condition, it will allow the possibility of disequilibrium where it will not hold. With disequilibrium in the commodity market, the economy will sometimes, but not at other times, have the aggregate demand for commodities equal to their supply. Interpreted in this way, Say's law would hardly merit the designation of a "law" since we do not give the equilibrium conditions for the money market, the bond market, the labor market or the foreign exchange markets – or for microeconomic markets such as those for apples – the designation of laws.

[14] The traditional classical approach followed this procedure and used the quantity theory to specify the price level. To do this, Say's law was supplemented by the equation $\cdot M\ P = m^{\mathrm{d}}\ P\ m_{\mathrm{y}} y = m_{\mathrm{y}} Y$, where y was set at full employment. Hence, Walras's law, Say's law and the quantity theory can be logically consistent with each

other, without such consistency implying their validity individually or as a set.

than being merely a negatively sloping curve.[15] There can be no meaningful IS–LM or IS–IRT analysis under such a shape of the IS curve. In fact, much of modern short-run macroeconomic theory would be ruled out in such a context.

Therefore, there are many reasons for rejecting the strong form of Say's law – that is, Say's law as an identity – for monetary economies.

Walras's law, Say's law and the dichotomy between the real and monetary sectors

A *dichotomy* (separation) is said to exist between the real (commodity) and financial (money and bonds) sectors if the demand and supply functions of the former are independent of nominal variables such as the price level, the inflation rate, the nominal interest rate and excess demand in the money and bond markets. In such a case, any shifts in the latter cannot change the values of the real variables under any circumstances.

Say's law alone implies that the demand for commodities always equals their supply, irrespective of the quantities of money, bonds and the price level. Therefore, the "real" system of the economy, which is concerned with the demand and supply of commodities and their relative prices, is independent of financial phenomena in the economy. There is thus a dichotomy between the real and financial sectors of such an economy. Such a dichotomy was also derived in Chapter 3 in the context of the general equilibrium version of the macroeconomic model, under the assumptions that the demand and supply functions were notional and possessed homogeneity of degree zero in all prices, rather than doing so in all prices *and* initial endowments. The present derivation of the dichotomy is related to the earlier one but is from a different perspective, with Say's law embodying within it the homogeneity of degree zero of the demand and supply functions of commodities in all prices. As Chapter 3 has shown, such a dichotomy does not exist in modern financial economies, which possess money and bonds.

Further, as Chapter 3 has argued, the *wealth and the real balance effects* show that a change in the price level changes the real financial wealth (composed of bonds and money balances) of the individual and changes his demand for real money balances and bonds, as well as his demand for commodities. Similar effects would occur if the price level was constant but the money supply was increased in such a way that the wealth of the individual and (the private part of) the economy changed. There is thus interaction between the commodities sector and monetary phenomena through the wealth and real balance effects, so that there does not exist a dichotomy between the real and the monetary sectors. Although these effects were explained in Chapters 3 and 14, we again do so briefly in the following.

The wealth effect

A change in the real value of wealth induces a *wealth effect* on the demand for commodities by households since this demand by individuals depends on their initial endowments (i.e. wealth). These initial endowments include households' and firms' holdings of money balances and

[15] Further, if the economy were only a commodities–money one, then, as argued earlier, Walras's law and Say's law together would imply that the demand for money will also always equal its supply, irrespective of the interest rate and income. Hence, in this case, the LM curve will also span the (r, y) space.

bonds, whose initial values are in nominal or dollar terms, so that a change in the price level of commodities changes the real value of the initial endowments.

In the debates between Keynesians and Keynes's critics in the 1940s and 1950s, the wealth effect is associated with A.C. Pigou, who argued that if aggregate demand was less than full-employment output, prices would fall and increase the real value of wealth. This would in turn increase consumption, thereby increasing aggregate demand and moving the economy to equilibrium at its full-employment level. The impact of the change in the price level on the real value of wealth, and of the latter on consumption, is known as the *Pigou effect*.[16] This effect operates as follows: the economy with deficient demand will continue to generate a price decrease, which will increase the real value of household wealth, which will increase consumption and increase aggregate demand until the demand deficiency is eliminated. While this argument is logical within the context of macroeconomic models, it relies heavily on the *ceteris paribus* condition, which does not normally hold in demand-deficient economies. In fact, in his later writings, Pigou argued that a fall in aggregate demand may not only cause a decline in the price level, it would also bring about simultaneous bankruptcies and deflation, with the consequence that real wealth may fall rather than increase, so that aggregate demand would fall rather than increase.[17] Therefore, the result of the original demand deficiency could more likely be a depression rather than a return to full employment. Hence, while the Pigou effect is an analytical ploy, its practical significance and validity as a device that returns a demand-deficient economy to full employment are doubtful.

At the general level, the wealth effect can occur because of changes in the real values of asset holdings, arising from changes in the nominal values of the assets or of their prices relative to the price level of commodities. The nominal values of the assets and their prices will change if the current or future expected interest rates change. Since changes in the price level can be accompanied by, or themselves induce, changes in interest rates, there can be both direct and indirect changes in wealth connected with changes in the money supply and the price level. Both these changes need to be incorporated in the short-run macroeconomic models. However, stable and predictable relationships between movements of the price level and movements in the prices of bonds, equities and physical assets have not been established. The postulates used for such relationships are at best gross simplifications and reflect a severe deficiency of knowledge on this topic.

The real balance effect

The real balance effect, associated with Don Patinkin's (1965) contributions from the 1940s to the 1960s, and already discussed in Chapters 3 and 14, is merely one element of the wealth effect and takes account of only those changes in real wealth that arise because of changes in the real value of money balances. This effect was defined in Chapter 3 as the change in the aggregate demand for commodities due to a change in the real value of the money balances held in the economy, with the latter due to a change in the price level or an exogenous change

[16] See also the discussion of the Pigou effect in Chapters 3 and 14.

[17] Chapter 14 pointed out that Pigou himself considered the effect named after him as a "mere toy" without empirical significance. We have used his argument: a fall in commodity demand will not only produce a price deflation (which increases demand) but, more significantly, it is also likely to cause or increase bankruptcies which will decrease production. The result is more likely to cause a depression than a return to full employment.

in the money stock. Such a change in real balances is part of the change in the individual's wealth due to the change in the price level. The real balance effect is one of the interactions that can occur between the commodity and the money markets and prevents money being neutral in the short run.[18] However, it only applies to outside money (i.e. M0), not inside money (i.e. bank deposits). It is not significant empirically.

Is Walras's law really a law? When might it not hold?

A *law* in economics has been interpreted in this chapter to be a statement that is true as an identity. It must hold under all states of the economy, from the most rudimentary to the most developed states, in recessions and in booms, in normal times and in chaotic conditions. For Walras's law to be a law, we should not be able to adduce any state of the economy, whether common or rare in the real world, which would not obey Walras's law. Walras's law was derived in Section 18.1 from the agents' budgetary constraints. But what would happen if there were also other constraints? Would the particular forms of the additional constraints vitiate Walras's law? The following subsection explores the implications of constraints imposed by the demand in the economy on the actual amounts of output that the firms are able to sell, the actual amounts of labor that households are able to sell and the actual amounts of commodities that households would buy.

Intuition: violation of Walras's law in recessions

The money and bond markets are so efficient in the modern financially developed economies that they adjust continually (every minute) while the financial markets are open, so that they can be taken to be continuously in equilibrium for analytical purposes. That is, we can take $E_m{}^{\mathrm{d}} = E_b{}^{\mathrm{d}} = 0$ *continuously*. Hence, (8) implies that:

$$E_c{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} = 0 \tag{14}$$

so that:

$$E_c{}^{\mathrm{d}} = -E_n{}^{\mathrm{d}} \tag{15}$$

Hence, there must exist positive excess demand for commodities whenever there is excess supply of labor. On a practical note, the evidence of (15) manifests itself in increases in unemployment during recessions, so that, in recessions, $E_n{}^{\mathrm{d}} < 0$. But recessions are also precisely the stage of the business cycle in which firms claim that the demand for their products has fallen and there is not enough demand for them to maintain their employment at the pre-recession levels, so that $E_c{}^{\mathrm{d}} < 0$. That is, the excess supply of labor and the excess supply of commodities occur concurrently in recessions, so that, in recessions, $E_c{}^{\mathrm{d}}\ E_n{}^{\mathrm{d}} < 0$. This evidence contradicts (14) and, therefore, Walras's law, thereby casting doubt not only

[18] The real balance effect is clearly relevant to the question of whether the economy can be in equilibrium below its full-employment level. This effect was an element in the disputes between the neoclassical economists and the Keynesians on the existence of an equilibrium below full employment, with the neoclassical school arguing that such a state would be one of disequilibrium; prices would decline and changes in aggregate demand and output would occur because of the real balance effect.

on its claim to be a law but also on its validity during recessions for economies with well-developed financial markets. This is a strong indictment of Walras's law and we will pursue its reasoning further in the following analysis.

The following two cases taken from the Keynesian paradigm (Clower, 1965/1969; Leijonhufvud, 1967, 1968, Ch. 2) examine the validity of Walras's law when there exists excess supply in the commodity and labor markets. Both cases assume an economy with efficient financial markets, so that there exists continuous equilibrium in the money and bond markets of the closed economy.

I. Unemployment in the labor market

For Leijonhufvud's arguments, start with the following implication of Walras's law when efficient financial markets ensure continuous equilibrium in the money and bond markets:

$$E^{\mathrm{d}}{}_{c} = -E^{\mathrm{d}}{}_{n} \tag{16}$$

Now assume that a shock to the economy produces involuntary unemployment. Since this means that there is excess supply of labor with $E^{\mathrm{s}}{}_{n} > 0$ (i.e. $E^{\mathrm{d}}{}_{n} < 0$), Walras's law implies that there must be positive excess demand in the commodity market. That is, firms' response to the rise in unemployment is to increase their production or raise their prices. This is counterfactual, as illustrated by economic analysts' usual interpretation of rising unemployment as an indicator of recession or forthcoming recession, so that their prediction becomes that of a cutback in production by firms.

Let us now follow our intuition and the usual analysis on the plausible and rational micro-economic behavior of households and firms. Since the unemployed workers corresponding to $E^{\mathrm{s}}{}_{n}$ do not receive any income, they, being rational, reduce their expenditures below those that they would have incurred had they been employed. Given this decrease in consumption expenditures, the commodity market will have an excess supply of commodities, where the supply is determined by what firms would wish to supply if they could sell all that they wanted to sell. In this eventuality, we would have:

$$E^{\mathrm{d}}{}_{c} < 0 \text{ and } E^{\mathrm{d}}{}_{n} < 0$$

so that:

$$E^{\mathrm{d}}{}_{c} + E^{\mathrm{d}}{}_{n} < 0 \tag{17}$$

Equation (17) contradicts the implication of Walras's law that we must have $E^{\mathrm{d}}{}_{c}\ E^{\mathrm{d}}{}_{n}\ 0$, when the money and bond markets are efficient and adjust continuously to equilibrium. In this scenario, which we have argued above as being quite plausible, the economy displays excess supply of labor without a corresponding excess demand in any market. This violates Walras's law. Hence, the law is not an identity and therefore not a law: it holds when there is equilibrium in all markets but may not hold when there is disequilibrium in the labor market.

What is required in this case for the *reinstatement* of Walras's law? What is required is the assumption that the unemployed workers continue to base their expenditures on the incomes they would have been entitled to receive *if* they had, in fact, been employed. But such posited behavior for households is implausible, irrational and violates the rational

Table 18.1 Walras's Law and the excess supply of labor

	E^{d}_{n}	E^{d}_{c}	E^{d}_{m}	E^{d}_{b}	$\Sigma_i E^{d}_{i}$
Walras's law:	$< 0 \Rightarrow$	$E^{d}_{c} = -E^{d}_{n}$ $E^{d}_{c} > 0$	0	0	0
Likely real-world and rational scenario in recessions:	$< 0 \Rightarrow$	$E^{d}_{c} < 0$ $E^{d}_{c} /= -E^{d}_{n}$	0	0	< 0

expectations hypothesis.[19] Alternatively, the reinstatement of Walras's law would require that the firms continue to pay laid-off workers their wages even though they have been laid off. This posited behavior for firms is also implausible, irrational, and violates the rational expectations hypothesis as applied to firms.

We summarize the preceding conclusions in Table 18.1, which has been compiled under the prior assumptions of efficient money and bond markets, so that they possess continuous zero excess demands, and involuntary unemployment.

In Table 18.1, while Walras's law derives E^{d}_{c} as an implication from E^{d}_{n}, the real-world scenario incorporates the *behavioral* relationship between E^{d}_{c} and E^{d}_{n} imposed by the optimization of its economic agents for the given structure of the economy and their expectations on wage income.

II. Excess supply in the commodity market

For the second case, this time from Patinkin (1965, Ch. 13), start with equilibrium in all markets, so that there is full-employment output. Again, assume that the money and bond markets are efficient, so that they are continuously in equilibrium. Now, let a fall in investment reduce aggregate demand and create excess supply (i.e. $E^{d}_{c} < 0$) in the commodity market. Faced with unsold output,[20] firms respond to their unsold output by reducing production and employment until the output of commodities becomes equal to their demand. While this adjustment by firms restores equilibrium in the commodity market through a reduction in their output to match the below-full-employment demand for commodities, it also reduces employment below full employment, so that a state of excess supply emerges in the labor market. Since, given the demand deficiency, there is no reason for firms to increase employment to its initial state of full employment, there continues to be excess supply of labor in the economy. Hence, the economy has excess supply in one market without a corresponding excess demand in any market. This violates Walras's law.

Table 18.2 highlights the arguments of this case on the conflict between Walras's law and the plausible real-world scenario.

[19] Rationality requires that economic agents base their decisions on all available information. In the present context, the workers know that they do not receive wage incomes while unemployed, so that they have to

decrease their consumption expenditures.

20 Under the perfect market hypothesis for commodity markets, the price will instantly adjust to eliminate excess supply. However, Walras's law does not assume perfect markets, even though the Walrasian general equilibrium model does so. For Walras's law to be an identity, it must hold whether markets are perfect or not.

Table 18.2 Walras's Law and the excess supply of commodities

	$E^d{}_c$	$E^d{}_n$	$E^d{}_m$	$E^d{}_b$	$\Sigma_i E^d{}_i$
Walras's law:	$< 0 \Rightarrow$	$= -E^d{}_c$	0	0	0
		$E^d{}_n > 0$			
Likely real-world and rational scenario in recessions:	$< 0 \Rightarrow$	$E^d{}_n < 0$	0	0	< 0
		$E^d{}_n /= -E^d{}_c$			

In Table 18.2, while Walras's law derives $E^d{}_n$ as an implication from $E^d{}_c$, the real-world scenario incorporates the behavioral relationship between $E^d{}_n$ and $E^d{}_c$ imposed by the optimization of its economic agents for the given structure of the economy and their expectations.

Walras's law under excess demand for commodities

For this section, we again assume that the money and bond markets are efficient and continuously in equilibrium. We now investigate the impact of a shock that produces an excess demand for commodities.

The impact of an excess demand for commodities

Suppose that a positive demand shock, such as to consumption, investment, fiscal deficits and imports, produces an excess demand for commodities. How does the economy usually respond to the increasing demand for its products? The plausible answer is that firms usually respond by adjusting their prices and output, with the output response normally preceding the price response, so that increases in aggregate demand first result in an increase in aggregate output, only later followed by inflation. The increase in production comes about through more intensive uses of both capital and labor, as in the implicit contract theory, as well as through increases in employment. Given the latter effect, the excess demand for commodities produces an excess demand (rising employment) for labor. Hence, the shock that had caused $E^d{}_c > 0$ led to $E^d{}_n > 0$, so that:

$$E^d{}_c + E^d{}_n > 0 \tag{18}$$

This scenario that $E^d{}_c > 0$ causes $E^d{}_n > 0$ is not only plausible, it is commonly observed: central banks and economic analysts use emerging evidence of increasing commodity demand (e.g. in factory orders) to predict a rise in production and a fall in unemployment.

Given the assumption that the money and bond markets are efficient and adjust continuously to equilibrium, (18) implies that $E^d{}_c + E^d{}_m + E^d{}_b + E^d{}_n > 0$. This contradicts Walras's law, which is that $E^d{}_c + E^d{}_m + E^d{}_b + E^d{}_n \equiv 0$.

Correction of Walras's law

The preceding analyses imply that the invalidity of Walras's law occurs not only in response to contractionary shocks to aggregate demand, which cause $E^{\mathrm{d}}{}_{c} < 0$ and recessions, but also to expansionary shocks to aggregate demand, which cause $E^d{}_c > 0$ and booms in the economy. In brief, in the economy, data showing $E^d{}_c < 0$ is a good basis for a prediction of rising unemployment ($E^d{}_n < 0$), while data showing $E^d{}_c > 0$ is a good basis for predicting falling unemployment ($E^d{}_n > 0$). Such a prediction pattern, commonly used by central banks and economic analysts, runs contrary to the prediction of Walras's law that data showing "high" aggregate demand (i.e. $E^d{}_c > 0$) would be a good basis for predicting falling unemployment ($E^d{}_n > 0$), while data showing depressed aggregate demand (i.e. $E^d{}_c < 0$) would be a good basis for predicting rising unemployment ($E^d{}_n < 0$).

Since Walras's law does not hold in such disequilibrium states,[21] it is really not a law (identity) but a statement about general equilibrium. Given its lack of validity for disequilibrium, its use in dynamic analysis can lead to misleading conclusions. It can, however, still be used for long-run analysis, since such analysis assumes equilibrium in all markets. Our arguments suggest that the correct statement of Walras's law is:

(i) In general equilibrium, $\sum_{k=1}^{K} E_k^d = 0$. However, this statement is trivial since, in general equilibrium, $E_k{}^d = 0_k$ for all k.

(ii) In disequilibrium, $\sum_{k=1} E_k^d$ can be positive (as in booms) or negative (as in recessions).

Notional demand and supply functions in the classical paradigm

For Walras's law to hold in the above two examples, there must be excess demand for output whenever there is excess supply of labor, and vice versa. Therefore, we need the actual clearance of all markets, or, if some of them are not in equilibrium, we need the hypothetical behavior pattern that *all agents continue to act (and expect) as if all markets cleared even when they do not.* Walras's law requires this assumption, which is unrealistic, to be an identity. As a consequence, the law requires that the demand and supply functions in (1) to (5) be *notional* functions, where a demand or supply function for a good *i* is said to be notional if economic agents act as if all *other* markets cleared. If this condition does not hold in practice, the operating functions would be *effective* functions, which are based on actual incomes and expenditures and do not assume that all other markets clear. They differ from the notional functions and Walras's law, as specified so far, would not apply to them. Hence, Walras's law would not be an identity for effective functions and strictly does not deserve the designation of a law.

Re-evaluating Walras's law

Fundamental causes of the failure of Walras's law

The preceding section shows that the assumption of equilibrium in the commodity and labor markets with disequilibrium in the bond and money markets does not lead to a violation of Walras's law. However, as shown in earlier sections, the assumption of continuous

[21] Note that with Walras's law as an identity, the arguments against it cannot be rejected by resort to special conditions, such as that of perfect markets. In any case, the assumption of perfect markets is not one of the assumptions of Walras's law.

equilibrium in the bond and money markets with disequilibrium in the commodities and labor markets does lead to such a violation. Why should the economy possess an asymmetry of this type? The reason lies in two positive relationships between commodities and labor. One of these comes from the supply side of the economy, that is, from the production function, which is of the type $y^s = y^s(n)$, so that an increase (decrease) in the production of commodities is accompanied by an increase (decrease) in employment. The other is from the demand side of the economy, for which the demand for commodities can be summarized by $y^d = y^d(n)$, where an increase in employment leads to higher incomes and higher commodity demand. These two relationships individually and together imply a strong positive relationship in practice between the output of commodities and the employment of labor, so that an excess positive demand for output must be accompanied by an excess positive demand for labor, and vice versa. Walras's law fails to take account of these fundamental behavioral relationships, whose implications run counter to the law and lead to its violation.

Irrationality of the behavioral assumptions behind Walras's law

To reiterate, given the starting assumption of continuous bond and money market clearance, Walras's law requires the clearance of both the labor and commodity markets. When they do not clear, it assumes that one of the following two things must occur:

1 Firms continue to produce the full-employment level of output if there is not enough demand for their products, or at least continue to pay workers the full wage even after laying them off.
2 Households behave as *if* the labor market had cleared at full employment, even if it means that unemployed workers spend incomes that they never received.

Both conditions are clearly irrational in the sense of not being profit/utility maximizing under the actual constraints that apply to firms and households, and do not hold empirically.

Reformulating Walras's law: the Clower and Drèze effective demand and supply functions

Clower effective functions

Clower (1965/1969) argued that in conditions where some markets fail to clear, the relevant demand and supply functions are not notional ones but must take account of disequilibrium in other markets. For example, if some workers are unable to sell their labor supply and are thereby involuntarily unemployed, their constraints on the purchases of commodities, money and bonds must take into account their resulting lack of labor income, so that their actual demands for these goods would be less than their notional demands. That is, their relevant demand functions incorporate the *spillover effects* of disequilibrium in the labor market. However, the relevant supply function of labor is still the notional one since the workers can still buy as much as they like of other goods.

As another example, if firms face deficient demand for commodities, i.e. they cannot sell all the output they wish to produce, they will have an effective demand for labor that is less than its notional demand. However, assuming that the firms do not face disequilibrium in markets other than for commodities, they would still operate with the notional supply function for commodities as their effective supply function. The demand and supply functions incorporating the impact of disequilibrium in *other* markets are called the *Clower effective demand and supply functions*. In microeconomic analysis, the effective demand/supply of an individual in market i is to be derived from the maximization of his utility function subject to his budget constraint *and* subject to the restrictions perceived by him in all other markets $k,\ k \neq i$. Similarly, the demand/supply of a firm in market i is derived from its profit maximization subject to its perceived constraints in markets $k,\ k \neq i$.

Clower (1965/1969) and Leijonhufvud (1967,1968) claimed legitimately that the Clower effective demand and supply functions – and not the notional ones – are the ones applicable to Keynesian analysis, with its emphasis on deviations from full employment, while the notional functions are applicable to classical analysis.[22]

Modification of Walras's law for Clower effective functions

Assuming that a disturbance has led to a fall in employment below the full employment level, we have a constraint in the form $n = n^{\mathrm{d}} \leq n^{\mathrm{s}}$, which modifies equation (1) to the inequality:

$$p_c c^{\mathrm{dh}} + p_c m^{\mathrm{dh}} + p_b b^{\mathrm{d}} + p_e e^{\mathrm{d}} \leq p_c \underline{c}^{\mathrm{s}} + \underline{M}^{\mathrm{h}} + p_b \underline{b}^{\mathrm{s}} + p_e \underline{e}^{\mathrm{s}} + Wn + \pi^{\mathrm{dis}} \tag{19}$$

where $Wn \leq Wn^{\mathrm{s}}$. Inequality (19), in combination with (2) to (5), yields the inequality:

$$E_c^{\mathrm{d\#}} + E_m^{\mathrm{d\#}} + E_b^{\mathrm{d\#}} + E_e^{\mathrm{d\#}} + E_n^{\mathrm{d}} \leq 0 \tag{20}$$

where the superscript # indicates effective functions,[23] Note that, in the case where Wn Wn$^{\mathrm{s}}$, these functions become notional ones and (20) changes to an equality. Clower argued that the proper statement of Walras's law is (20), that is, the sum of all actual excess demands in the economy is *non-positive*. In this statement of Walras's law, in the context of our assumed involuntary unemployment, the excess demands for goods other than labor are effective ones while the demand for labor is notional. Correspondingly, note that if the perceived constraint by firms was that of deficient demand for commodities – that is, $y^{\mathrm{d}} < y^{\mathrm{s}}$ – the excess demand for commodities would be notional while the excess demands for the other goods would be effective.

A consequence of (20) is that it cannot be used to derive the statement that equilibrium in K 1 markets implies equilibrium in the K th one also, since the $K-1$ markets with demand equal to supply could be those with effective functions while the K th one could be the one with the disequilibrium, with its disequilibrium being the inequality of its notional demand and supply.

Drèze effective functions and Walras's law

In the context of involuntary unemployment, Drèze (1975) proposed a reinstatement of Walras's law by equating the supply of labor to the effective supply set by the smaller demand for labor. That is, set $n^{\mathrm{s}}{}_{\mathrm{D}} = \bar{n}^{\mathrm{d}}$, where $n^{\mathrm{s}}{}_{\mathrm{D}}$ is the Drèze supply of labor. Its purpose

[22] The two are identical if equilibrium exists in all markets.

[23] For such functions, a market k can have effective clearance (i.e. $E_k{}^{\mathrm{d\#}} = 0$), with or without notional market clearance (i.e. $E_k{}^{\mathrm{d}} = 0$).

is to impose *all* the constraints that operate in the economy. In the general case, this results in the Clower demand/supply functions for all markets *i* in which there are no perceived constraints, but with the Drèze demand/supply functions set by the constraints themselves in the constrained *k* markets. Equations (1) to (5), with demands and supplies redefined in this manner, again imply Walras's law, but in a Drèze effective – and not a notional – sense. Excess demands would also be redefined in a corresponding manner and the sum of all such (Drèze) excess demands would again equal zero.

Note that we can speak of equilibrium in an unconstrained market in the sense of the equality between the Clower demand and supply in that market, and determine its price from such equality. However, we cannot consider the equality of Drèze demand and supply in a constrained market as representing equilibrium in it, or use such equality as a basis for determining the price in it. Further, the Drèze constraint ${n^s}_{\mathrm{D}} = \bar{n}^d$ is strange: it specifies the supply of labor by its demand. There is no intuitive justification for it, as there is for the Clower analysis. Note that both Clower's reformulation of Walras's law as an inequality and Drèze's reinstatement of it in terms of Drèze functions limit the usefulness of this law.

Implications of the invalidity of Walras's law for monetary policy

What has the preceding discussion on the invalidity of Walras's law in disequilibrium got to do with monetary theory and policy, which is the subject matter of this book?

This chapter has shown that, in disequilibrium with a violation of Walras's law, a shock that produces an aggregate demand deficiency (excess supply) of commodities can be accompanied by an excess supply (unemployment) of labor, and often is. The mechanisms that were adduced to restore such an economy to equilibrium have been the real balance and the Pigou effect. These depend on the demand deficiency producing a fall in the price level, which increases the value of real balances and bonds, causing the wealth effect on consumption to increase aggregate demand in the economy. This process will eventually eliminate the demand deficiency. However, even if such mechanisms were to operate as posited, this process would be extremely slow and may take decades to eliminate a serious demand deficiency. Such a delay is an invitation to the central bank to try to reduce its duration to a much shorter period. The central bank can do so through expansionary monetary policies. These can be reductions in the interest rate or increases in the money supply, but usually both. Such responses to demand deficiency are, in fact, embodied in the Taylor-type rules of monetary policy formulation.

The implication of the invalidity of Walras's law in disequilibrium for a realistic monetary policy is that the relevant dynamic analyses of the movements in output and the price level produced by the economy should be based on the effective excess demand functions, not the notional ones. This is, in fact, the actual practice followed by central banks and economic analysts in predicting future movements in output and prices and the need for an active monetary policy. To illustrate, the output gap in Taylor rules is the deviation of effective (not notional) output from its full-employment level.

Conclusions

Walras's law is a core underlying relationship in the specification of macroeconomic models and is used to eliminate the explicit treatment of one of the markets in them. By convention in both the classical and Keynesian types of macroeconomic models, the market thus rendered implicit in the analysis is usually the bond market, though it could be any of the others.

Say's law, properly interpreted, is an identity between the supply and the demand for commodities. However, it is not valid for a economy with money and bonds, since such an economy allows substitution between financial assets and commodities – and thereby possesses interaction, at least in the disequilibrium states, between the monetary and commodity sectors.

Walras's law and Say's law imply a dichotomy between the real and monetary sectors and the neutrality of money. The dichotomy is definitely not valid in a monetary economy. The validity of the neutrality of money in the short run depends upon the structure of the economy, since it requires wage and price flexibility, continuous market clearance and the absence of the real balance effect in equilibrium. Few economies meet these stringent conditions in the short run.

While Walras's law is an identity in the context of the *notional* demand and supply functions, it becomes merely an inequality with the Clower effective demand and supply functions. The use of Drèze functions reinstates its equality but at the cost of its usefulness. Keynesian analyses are usually based on Clower effective functions.

The Pigou and real balance effects specify, respectively, the impact of changes in financial wealth and real balances on the aggregate demand for commodities. They are, therefore, among the elements interconnecting the commodity and financial markets and played a critical role historically in discussions on the dichotomy between the sectors. These effects clearly exist in the short run but have limited empirical relevance. The Pigou effect becomes even more doubtful if the impact of a price deflation on interest rates and the insolvency of debt-laden firms are taken into consideration.

Summary of critical conclusions

- Walras's law is based on the budget constraints of the various economic agents in the economy and is perhaps the closest we can get to an identity in economics.
- Say's law does not apply in monetary economies.
- Walras's law can be used to derive the excess demand function for bonds, which is an asset in the macroeconomic framework even though its demand and supply functions are often left unspecified in the standard IS–LM analysis.
- The real balance and Pigou effects are important theoretical links between the money and the commodity markets, but their empirical importance in the modern economy is limited.
- Keynesian analyses of deficient demand and involuntary unemployment use effective demand and supply functions, and modify Walras's law to an inequality.
- Clower and Drèze demand and supply functions are more relevant than notional ones in

Review and discussion questions

1. Walras's law is derived from the budget constraints of all the economic agents in the economy. Can Say's law be similarly derived from budget constraints? Use the relevant constraints and specify the additional assumptions needed for this derivation. Assess the validity of these assumptions.

2. What are the implications for monetary policy if both Walras's law and Say's law are imposed on the IS–LM model? Assess the likely validity of these implications. If they do not seem to be valid, which of these two laws should be discarded? Derive the implications for monetary policy of imposing the remaining law on the IS–LMmodel.
3. Do the modern classical or/and new classical schools effectively reinstate Say's law as one of their component doctrines? Is so, should they state it explicitly? Discuss.
4. Derive the implications of Walras's law and Say's law together for the determinacy of absolute and relative prices in a commodities–money (no bonds or labor) economy. What role does the real balance effect (in the short run and the long run) play in this determination?
5. Outline your understanding of households' and firms' most likely responses to a fall in the aggregate demand for commodities. Does its perceived duration matter? If a downturn in the economy leads to a fall in demand that is perceived to be significant in magnitude and duration, discuss whether Walras's law will continue to hold. If it does not, what happens to the excess demand for commodities and for bonds in the IS–LMmodel?

   Does Walras's law hold in recessions?
6. In terms of your understanding and beliefs about the functioning of your economy, which of the four markets (commodities, money, bonds and labor) clear on a daily, weekly and monthly basis? Which may not do so, at least within thirty days of a disturbance? Within 6 months? Within a year?

   If some of the markets do so while others do not do so, does this support Leijonhufvud's (1967,1968) and Clower's (1965/1969) contention that Walras's law is not an identity when there is disequilibrium in some markets?
7. For the magnitudes of the relevant variables in any of the past five years in your economy, try to assess the importance of the real balance effect for a 5 percent fall in the price level.
8. What was the dichotomy between the real and the monetary sectors in the traditional classical approach to macroeconomics? How would such a dichotomy arise in aWalrasian general equilibrium system?

   What was the contribution of Patinkin's real balance effect to this debate?
9. Does the dichotomy between the real and the monetary sectors hold in the modern classical approach? Is this dichotomy valid or not for the modern financially developed economy?
10. What were Keynes's arguments in his attacks on Say's law and the traditional classical dichotomy? In retrospect, and especially in the light of the reversion (counter- reformation!) of macroeconomics to the (modern version of the) classical model, evaluate the success of these attacks.
11. Keynes argued that an economy could be in equilibrium with a substantial amount of involuntary unemployment, but other economists disagreed and argued that a state in which an important market does not clear is one of disequilibrium. Explain the notions of equilibrium and disequilibrium involved, Keynes's justification for his position, and his opponents' justification for theirs. Does the existence of the real balance effect or the wealth effect refute Keynes's position?
12. Discuss: "The real balance effect provides a possible dynamic explanation of the adjustments in the economy in going from one equilibrium to another and is an effective answer to the assertion of a Keynesian under-full-employment equilibrium.

However, it is not really necessary for the comparative static propositions of the quantity theory or of neoclassical economics."

13. Discuss: "The derivation using Walras's law of the excess demand function for bonds from those of other markets provides insights into its properties and also shows some clearly invalid assumptions usually made for the excess demand and supply functions for the other markets". Illustrate with examples from IS–LM analysis.
14. Robert Clower argued that "either Walras's law is incompatible with Keynesian economics, or Keynes had nothing fundamentally new to add to orthodox economic theory." Is Walras's law incompatible with the different Keynesian and neoKeynesian models as they have evolved?
15. Define the notional, Clower and Drèze demand and supply functions. Prove whether or not Walras's law applies to each of these three types of functions and in what sense it does so.

References

Baumol, W.J. "Say's law." *Journal of Economic Perspectives*, 13, 1999, pp. 195–204.

Clower, R. "The Keynesian counter-revolution: a theoretical appraisal." In F.H. Hahn and

F.P.R. Brechling, eds, *The Theory of Interest Rates*. London: Macmillan, 1965. Also in R.W. Clower, ed., *Monetary Theory, Selected Readings*. London: Penguin, 1969.

Drèze, J.H. "Existence of an exchange equilibrium under price rigidities." *International Economic Review*, 16, 1975, pp. 301–20.

Lange, O. "Say's law: a restatement and criticism." *Studies in Mathematical Economics and Econometrics: In memory of Henry Schultz*. Chicago: University of Chicago Press, 1942.

Leijonhufvud, A. "Keynes and the Keynesians." *American Economic Review Papers and Proceedings*, 57, 1967, pp. 401–10.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press, 1968.

Patinkin, D. *Money, Interest and Prices*, 2nd edn. New York: Harper & Row, 1965. Shaller, D.R. "Working capital finance considerations in national income theory." *American Economic*

Review, 73, 1983, pp. 156–65.

Part VI

The rates of interest in the economy

19 The macroeconomic theory of the rate of interest

The rate of interest is one of the endogenous variables in the Keynesian and classical models, so that its analysis is properly conducted as part of a complete version of those models, which were presented in Chapters 13 to 15.

This chapter singles out the competing views on the determination of the rate of interest and focuses on their differences and validity. It also highlights the very important difference between the comparative static and the dynamic determination of the rate of interest.

Key concepts introduced in this chapter

- ♦ Fisher equation of the nominal rate of interest
- ♦ Stocks versus flows of funds
- ♦ Loanable funds theory
- ♦ Liquidity preference theory
- ♦ Excess demand function for bonds
- ♦ Dynamics of interest rate determination
- ♦ Neutrality of money and inflation for the real rate of interest

As explained in Chapters 13 to 15, macroeconomic and monetary analysis until recent decades, and usually even now, has assumed only two distinctive financial assets, money and non-monetary financial assets, and allocated the terms "bonds," "credit" and "loanable funds" as synonyms for the latter. Traditional classical (pre-1936) economists had preferred the term "loanable funds," while modern analysis, as in Chapter 13, prefers the term "bonds" to designate all non-monetary assets. This chapter will use these terms interchangeably, as in Chapters 13 to 15, rather than following the distinctive macroeconomic analysis of Chapter 16, which had two non-monetary financial assets (bonds and credit/loans).

The interest rate is the return on bonds (all non-monetary financial assets), but we have not so far in this book specified explicitly the demand and supply, or the excess demand, functions for bonds. There are two ways of doing so. One is to derive the excess demand for bonds, using Walras's law, from the demand and supply of the other three goods (commodities, money and labor) in the macroeconomic model. The other method is to derive the bond demand and supply functions directly from the behavior of economic agents. We present both of these procedures in this chapter.

Bonds in this chapter comprise a single homogeneous, non-monetary financial asset.

Further, to get around issues raised by maturing bonds and to establish a simple relationship between the nominal bond price p_b and the nominal interest rate R, the (homogeneous) bond is assumed to be a consol (perpetuity), which promises a constant coupon payment of $1 in perpetuity. For this consol, $p_b = 1/R$.

This chapter studies the comparative static and dynamic determination of the macroeconomic interest rate in the closed economy. In terms of the heritage of ideas, the theories that deal specifically with the determination of this interest rate are the traditional classical loanable funds theory and the Keynesian liquidity preference theory. The loanable funds theory asserts that the bond market determines the interest rate, whereas the liquidity preference theory asserts that the money market does so. The coverage of these theories and their validity is an important part of this chapter.

Given the common or underlying macroeconomic rate of interest as determined in this chapter, the next chapter will examine the time aspects of the interaction among the various interest rates in the economy. The main focus of that chapter will be on the term structure of interest rates.

Section 19.1 reviews the Fisher relationship between the real and nominal interest rates and the impact of inflation on interest rates. Sections 19.2 to 19.6 examine the implications of Walras's law for the determination of the interest rate and the derivation of the excess demand for bonds. Section 19.7 looks at the modern and the historical versions of the loanable funds theory of the rate of interest. Section 19.8 presents the Keynesian liquidity preference theory of interest. Section 19.9 compares the loanable funds and liquidity preference theories in the comparative statics context and shows that the two yield identical comparative static implications for the rate of interest, while their dynamic analyses give different implications, so that a choice has to be made between them. Section 19.10 discusses the neutrality of money for the real interest rate. Sections 19.12 and 19.13 present empirical findings on the Fisher equation, and on the loanable funds and liquidity preference theories.

Nominal and real rates of interest

The Fisher equation on the interest rate

As explained in Chapter 2, the Fisher equation is:

$$(1 + r^e) = (1 + R)/(1 + \pi^e) \qquad (1)$$

where R is the nominal interest rate, r is the real interest rate, r^e is the expected real interest rate and π^e is the expected inflation rate. If there exist both real bonds (i.e. promising a real rate of return r per period) and nominal bonds (i.e. promising a nominal rate of return R per period), the relationship between them in perfect markets would be:

$$(1 + R) = (1 + r)(1 + \pi^e) \qquad (1^J)$$

At low values of r^e and π^e, $r^e\pi^e \rightarrow 0$, so that the Fisher equation is often simplified to:

$$r^e = R - \pi^e \qquad (1^{JJ})$$

Mundell–Tobin effect of expected inflation on the real interest rate

In (1), on the interaction between the real interest rate and the expected inflation rate in the context of an exogenous money supply, Mundell (1963) argued, in the context of the

IS–LM analysis with an exogenous money supply, that an increase in the expected inflation rate would cause a reduction in the demand for real balances, which would lower the real interest rate. This came to be labeled the *Mundell effect*. Tobin (1965) argued that, in a general macroeconomic model with a variable physical stock, a reduction in the demand for real balances due to the Mundell effect will increase the demand for real capital,[1] so that as capital accumulates, its productivity falls, which drags down the real interest rate. This is known as the *Tobin effect*.[2] The combined impact of higher expected inflation on the real interest rate is often called the Mundell–Tobin effect.

Impact of high and persistent money growth on the nominal interest rate

Note the impact of changes in the money supply on the nominal interest rate through its impact on the real interest rate and the expected rate of inflation. An increase in the money supply lowers the real rate (the Mundell effect) in the IS–LM analysis through a rightward shift of the LM curve, but it also causes inflation which, through the formation of expectations, creates expected inflation and, through the Fisher equation, raises the nominal rate. At very low rates of expected inflation, the net effect of money creation is often to lower the nominal (and real) interest rates, at least for some time until expected inflation catches up to actual inflation. Over time, high and persistent rates of inflation become expected rates and are invariably accompanied by high nominal rates.

Application of Walras's law in the IS–LM models: the excess demand for bonds

Walras's law

The derivation of Walras's law was presented in Chapter 18. For the compact four-good (commodities, money, bonds and labor) macroeconomic model, Walras's law is the identity that *the sum of the nominal excess demands for all goods in the closed economy must be zero*. That is,

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{2}$$

where $E_k{}^{\mathrm{d}}$ is the excess *nominal* demand for the *k*th good, $k = c, m, b, n$. c is consumption, m is real balances, b is bonds and n is labor. Spelled out, (2) is the identity that:

$$P(c^{\mathrm{d}} - c^{\mathrm{s}}) + P \cdot (m^{\mathrm{d}} - m^{\mathrm{s}}) + p_b(b^{\mathrm{d}} - b^{\mathrm{s}}) + W(n^{\mathrm{d}} - n^{\mathrm{s}}) \equiv 0 \tag{3}$$

where the superscripts d and s stand for demand and supply respectively, and:

P = price of commodities (the price level)
p_b = price of bonds
W = nominal wage rate (rental price per period of labor)
c = quantity of commodities

[1] With labor, money balances and capital in the production function, fairly standard assumptions imply that a decrease in money balances would increase the demand for both labor and capital.

[2] The empirical significance of the Tobin effect is probably negligible in a short-term, as well as a long-term,

context. It would require the variability of physical capital in short-run models.

m = real money balances (= M/P)
M = nominal money balances
b = quantity of bonds
n = number of workers (labor).

Implications of Walras's law for a specific market

Equation (3) implies that:

$$E_K^{\mathrm{d}} \equiv -\sum_{k=1}^{K-1} E_k^{\mathrm{d}} \tag{4}$$

where:

$E_k{}^{\mathrm{d}}$ = excess nominal demand in the kth market
$E_K{}^{\mathrm{d}}$ = excess nominal demand in the K th market.

Equation (4) allows the excess nominal demand in the K th market to be derived from the excess nominal demands in the other (K–1) markets. Note that we can arbitrarily designate the market for any specific good as the K th market. (4) implies that:

$$\text{If } \sum_{k=1}^{K-1} E_k^{\mathrm{d}} = 0, \text{ then } E_K{}^{\mathrm{d}} = 0 \tag{5}$$

which implies the *conditional* statement that *if* there exists equilibrium in $K - 1$ markets, *then* there would also be equilibrium in the *K*th market. For *comparative static analysis,* (4) allows the analysis to dispense with the explicit treatment of one of the markets in the economy. However, such an omitted market continues to exist and to function but its treatment is pushed into the implicit state. Which market is omitted depends on custom, convenience and the purpose at hand.

Implications of Walras's law for the solution of equilibrium prices

As is clear from (3), the four-good economy has only three prices in it. These are the price level P as the price of commodities, the bond price p_b (or its alter ego, the interest rate) and the nominal wage rate W as the (rental) price of labor. Since Walras's law implies that equilibrium in three markets also ensures equilibrium in the fourth one, we can find the three equilibrium prices by solving the equilibrium conditions for any of the three markets. The calculated equilibrium prices will be invariant to the selection of the three markets for the model being solved.

Walras's law and different groupings of markets in monetary and macroeconomics

The preceding analysis implies that, without changing the equilibrium solution of the three prices, the explicit statement of our macroeconomic model can include only:

(I) money, bond and labor markets;
(II) commodity, money and labor markets;

(III) commodity, money and bond markets;
(IV) commodity, labor and bond markets.[3]

The selection among these choices depends on custom, convenience and the purpose at hand.

The traditional classical (pre-Keynes) economists chose grouping I. They specified the quantity theory for the money market, the loanable funds theory for the bond market and the labor market for the determination of employment. They did not specify a theory for the aggregate demand and supply of commodities and did not explicitly include the commodity market in their analyses.[4] By comparison, following Keynes, the Keynesian school chooses grouping II and specifies the commodity market, the money market and the labor market, but leaves out an explicit analysis of the bond market. The neoclassical and modern classical schools also follow this pattern. However, as we have shown above, as long as the structural equations of the macroeconomic model are the same, Walras's law ensures that the equilibrium values of the three endogenous prices, as well as other real variables such as prices, output, investment, consumption, etc., will be identical among the models. Hence, for comparative static analysis (which solves or compares only the general equilibrium values), the selection of the three markets for explicit analysis is immaterial for the representation of the economy.

An assumption commonly made in modern classical economics is that the labor market is continuously in equilibrium. In this case, Walras's law implies that

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} = 0 \tag{6}$$

However, empirically, the underlying assumption of (6) that the labor market continuously clears – while the commodity, money and bond markets may not do so – is highly questionable.

Derivation of the general excess demand function for bonds

Chapter 13 specified for the open economy the general form of the demand function for commodities by the IS equation as:

$$y^{\mathrm{d}} = y^{\mathrm{d}}(r, P; \lambda) \tag{7}$$

where λ represents the fiscal policy variable, especially the fiscal deficit. The analysis in Chapter 14 of the labor market for the classical paradigm without uncertainty specifies the equilibrium condition for the labor market as:

$$n^{\mathrm{d}}(w) = n^{\mathrm{s}}(w) \tag{8}$$

which determines the equilibrium real wage *w*, which when substituted in the labor demand function determines the full-employment level *n* as n^{f}. Substituting the latter in the production function:

$$y = y(n) \tag{9}$$

[3] Note that it is rare to find macroeconomic analysis based on groupings III and IV.

[4] This is not surprising since the analysis did not include the concept of the consumption function and the multiplier.

determines the supply of commodities y^{s} as y^{f} . y^{f} depends on the supply of labor and technology but is independent of the other variables of the model, so that the classical paradigm's supply function in the absence of uncertainty[5] is:

$$y^{\mathrm{s}} = y^{\mathrm{f}} \tag{10}$$

However, the Keynesian paradigm assumes commodity market imperfections and specifies the commodity supply function by a price/quantity adjustment function, two versions of which are the Phillips curve and the new Keynesian Phillips curve (see Chapter 15). For a comparative statics model, specify the general form of the Keynesian output supply function as:

$$y^{\mathrm{s}} = y(P) \tag{11}$$

Therefore, given the commodity demand function (7) and irrespective of whether we use the commodity supply function of the classical or the Keynesian paradigm, the excess demand $e_c{}^{\mathrm{d}}$ $(= y^{\mathrm{d}} - y^{\mathrm{s}})$ function for commodities has the general form:

$$E_c{}^{\mathrm{d}} = P \cdot e_c{}^{\mathrm{d}}(r, P; \lambda) \tag{12}$$

where $e_c{}^{\mathrm{d}}$ is the commodity excess demand in real terms and $E_c{}^{\mathrm{d}}$ is its nominal value. λ represents the fiscal policy variables.

From Chapter 13, the excess demand function for real balances $e_m{}^{\mathrm{d}}$ is of the general form:

$$E_m{}^{\mathrm{d}} = P \cdot e_m{}^{\mathrm{d}}(y, R, P; \theta \cdot \mathrm{M0}, FW_0) \tag{13}$$

where θ .M0 is the money supply, M0 is the monetary base and θ is the multiplier from the monetary base to the money supply. FW_0 is the initial amount of financial wealth.[6]

The analysis of the labor market in Chapter 14 implies the excess demand functions for labor $e_n{}^{\mathrm{d}}$ as:

$$E_n{}^{\mathrm{d}} = W \cdot e_n{}^{\mathrm{d}}(w) \tag{14}$$

Hence, by Walras's law as stated in equation (4), the excess real demand function $e_b{}^{\mathrm{d}}$ for bonds is:

$$p_b \cdot e_b{}^{\mathrm{d}} = -P.e_c{}^{\mathrm{d}}(r, P; \lambda) - P \cdot e_m{}^{\mathrm{d}}(y, R, P; \theta \cdot \mathrm{M0}, FW_0) - W.e_n{}^{\mathrm{d}}(w) \tag{15}$$

so that the general form of the excess nominal demand ($E_b{}^{\mathrm{d}}$) for bondsis:

$$E_b{}^{\mathrm{d}} = E_b{}^{\mathrm{d}}(R, P, w; \lambda, \theta.\mathrm{M0}, FW_0, \pi^{\mathrm{e}}) \tag{16}$$

[5] The impact of uncertainty on y^{s} generates a short-run aggregate supply curve, as shown in Chapter 14.

[6] *FW* is generally not specified as an argument of the money demand function. However, omitting it there implies that the excess demand for bonds would not depend on financial wealth, which would be patently

invalid and analytically undesirable.

which omits *y* since *y* equals *w.n* in this model. π^e appears as an argument because of the Fisher equation. Alternatively, *w* can be omitted and replaced by *y*. Doing so would mean writing the excess nominal demand function for bonds as:

$$E_b{}^{d} = E_b{}^{d}(R, P, y; \lambda, \theta.\mathrm{M0}, FW_0, \pi^e) \tag{17}$$

Intuition: the demand and supply of bonds and interest rate determination

Since *R* is the nominal return on bonds, its equilibrium value is determined by:

$$b^d(R, \ldots) = b^s(R, \ldots) \tag{18}$$

where *b* is the number of (homogeneous) bonds/consols. We have assumed that the demand and supply of bonds depend on the nominal interest rate, among other variables. Both the demand for and supply of bonds have a flow and a stock dimension.

Flows versus stocks

In terms of *flows over a specified period of time*, the demand for bonds corresponds to the amount of (loanable) funds flowing or coming onto the market for lending at the various rates of interest. Similarly, the supply of bonds corresponds to the demand for (loanable) funds from those wanting to borrow funds during the period. However, the flow of funds that becomes available for loans over the current period is only a small fraction of the total amount of credit outstanding in the economy. This total amount is like a reservoir and is the *stock* of loanable funds. The stock of loanable funds supplied at any point in time consists of all outstanding loans plus the net additional flow supply of loanable funds, specified for each rate of interest. In stock terms, the demand for credit is similarly the total amount already borrowed plus the net additional amounts that the borrowers wish to borrow at each rate of interest. In modern economies, a major part of this demand often comes from the existing public debt.

In markets with long-term contracts, some of the borrowers and lenders are already committed to loans made at rates prevailing in the past. In such a case, the proper market for determining the current rate of interest is that in terms of flows: the flow market is the actual operating market for bonds in any given period, with borrowers (sellers of bonds) entering it to borrow and lenders (buyers of bonds) entering it to lend funds. However, note that the pre-existing stock of bonds does exert a strong background influence on the flow demands and supplies since parts of this stock of bonds will be expected by borrowers and lenders to mature sooner or later and, over time, become flows available for renegotiation.

The *flow supply of funds* can be interpreted as that part of the stock that has come up for renegotiation plus the additions being made currently. The net *new* supply of funds to the credit market in any period *t* comes from two sources:

(i) Current (private) saving in the economy.
(ii) Excess supply of money made available for loans, with the excess supply resulting from changes in the public’s desired balances or in the supply of money. The supply of money depends on the monetary base and the inside money created by financial intermediaries.

The overall supply of funds in period *t* is the net new supply from the above two sources plus:

(iii) Funds becoming available from loans that have matured in period *t*.

The *flow demand* for loans is from net new borrowers and those who wish to renew existing loans. The net *new* demand for loans comes from:

(iv) Current investment in the economy.
(v) Bond-financed government deficits.[7]

The flow demand for loans in period *t* is from (iv) and (v), plus:

(v) Demand for credit from those whose loans have matured.

Assuming (iii) and (vi) to be equal, the loanable funds theory in flow terms specifies the real demand (f^{d}) and supply (f^{s}) functions of loanable funds as:

$$f^{\mathrm{s}} = s(r, \ldots) + (\theta.\mathrm{M0}^{\mathrm{s}}/P - m^{\mathrm{d}}(R, \ldots)) \tag{19}$$

$$f^{\mathrm{d}} = i(r, \ldots) + (g - t) \tag{20}$$

where:

f^{s} = real flow supply of loanable funds (demand for bonds)
f^{d} = real flow demand for loanable funds (supply of bonds)
s = real saving
i = real investment
g = real government expenditures
t = real government revenues
$\mathrm{M0}^{\mathrm{s}}$ = supply of the nominal monetary base
θ = monetary base to money supply multiplier ($= \partial M^{\mathrm{s}}/\partial \mathrm{M0}$)
m^{d} = demand for real balances
P = price level.

We have assumed that the government deficit (g_t) is wholly bond-financed and that r and R are related by the Fisher equation. In partial equilibrium analysis, the equilibrium value of the market rate of interest is determined by:

$$s(r, \ldots) + (\theta.\mathrm{M0}^{\mathrm{s}}/P - m^{\mathrm{d}}(R, \ldots)) = i(r, \ldots) + (g - t) \tag{21}$$

Note that the left side of (21) represents the demand for bonds and the right side represents the supply of bonds. (21) is the statement that the interest rate is determined by the equilibrium in the flow part of the bond market.

Long-run determination of the interest rate

Equation (21) specifies the determination of the short-run interest rate and shows that, although the interest rate is determined by the excess demand for loanable funds, it is

[7] This category would be negative for a budget surplus.

not independent of the excess demand for money:[8] excess money demand raises the interest rate and excess money supply lowers it. However, money supply and demand enter the determination of the interest rate *only if* there is disequilibrium in the money market.

In the long run (general equilibrium) the money market would be in equilibrium, so that the excess money demand term on the left side of (21) is zero. Hence, the long-run version of the bond market analysis becomes:

$$s(r,\ldots) = i(r,\ldots) + (g - t) \tag{22}$$

or:

$$s^{n}(r,\ldots) = i(r,\ldots)$$

where s^n is national saving (s $(t$ $g)$). In the context of the closed economy, (22) is also the statement of equilibrium ~~in the~~ commodity sector of the economy.

Intuition: dynamic determination of the interest rate

We first illustrate the nature of our further arguments by starting with an illustration from the commodity markets. Suppose that equilibrium does not hold in a particular market, say for peanuts. Then the excess demand for peanuts would be a function of the peanut price and of the prices of its substitutes and complements, and the price of peanuts will change in response to the excess demand for peanuts. Further, in general, the larger the excess demand, the faster will be the price change. This adjustment in the price of peanuts is, however, not *directly* influenced by the existence and extent of disequilibrium in the market for other products,[9] even though all markets are related by Walras's law. If there is disequilibrium in a market for a close substitute for peanuts, say potato chips, this disequilibrium will flow into the demand for peanuts, creating disequilibrium in the market for peanuts and influencing their price. But this change in the price of peanuts is not a direct affect of the disequilibrium in the market for chips on the price of peanuts but rather an indirect effect occurring from the spillover, because of substitution, into the demand for peanuts, and depends upon the small or large amount of that spillover. That is, the price of a good responds in a dynamic context directly to the excess demand for that good, with the state of excess demands for other goods being either irrelevant or indirectly relevant in so far as they first affect the excess demand for the good in question.

The rate of interest is a price. But what is the good with which it should be identified in a dynamic context? One approach to this is the liquidity preference concept, explained in detail later, which would identify the interest rate with the good "money," thereby making dynamic changes in the interest rate R a function of the excess demand for money $E_{mt}{}^{d}$, so that $\partial R_t/\partial t = f\,(E_{mt}{}^{d})$. The alternative approach is that of loanable funds, which would define

[8] Note that the labor market does not appear explicitly in the loanable funds equations (15) to (21), but the determination of incomes in the labor market is clearly a determinant of saving in the economy, so that the labor market is implicitly included in the determination of the loanable funds interest rate.

[9] That is, the price of peanuts is not directly a function of the excess demands for other commodities, including, say, almonds, apples or chairs; though the excess demand for peanuts could be indirectly made, through appropriate substitutions, a function of the excess demands for the other commodities.

the relevant good as bonds, thereby making the dynamic changes in the interest rate a function of the excess demand for bonds $E_{bt}{}^{d}$, so that $\partial R_t / \partial t \underset{=}{\_} f\,(E_{bt}{}^{d})$. These two approaches yield different rates of change in the interest rate.

At a practical level, the dispute between the traditional classical and the Keynesian theories of the rate of interest comes down to which approach will do better empirically in a *dynamic* context. However, there is no generally accepted empirical evidence on this issue, so we should not ignore our intuition on it. Our intuition on the *operational* markets in the economy has already been specified in Chapter 18 and is along the following lines.

In a monetary economy, commodities are always bought and sold at a price against the specific good that is labeled money. There are, therefore, operational markets for commodities (or an operational market for "the commodity" in a single commodity model), so that the equilibrating variables are commodity prices (and the average price level) in commodity markets. Loans are made, again always in a specific good that is money, and the interest rate is agreed between borrowers (sellers of bonds) and lenders (buyers of bonds). It is then the "price" of loans in the bond market. There is, therefore, an operational market for loanable funds (bonds), with the interest rate as the equilibrating variable.

Note that there is no real-world market where money is always bought and sold against *one* specific good. If individuals want to run down their money balances, they can do so either by buying goods in the commodities markets or by making loans in the credit market, with different individuals making this choice in different proportions. These arguments lead to our hypothesis: *the economists' market for money balances is an analytical construct without an operational real world counterpart*.[10] The hypothetical market for money arises only because of Walras's law, a comparative statics concept, and is, as it were, a composite reflection of the other markets. It can be used only for comparative static analysis but not for dynamic analysis to determine the price level or the interest rate in a dynamic disequilibrium context. To conclude, in a dynamic context, the proper analysis of the interest rate is through bond market analysis, as in the loanable funds theory, and not through money market analysis, as in the liquidity preference theory.

The above arguments are buttressed by the buffer stock role of money. As shown in Chapter 6 above, the buffer stock concept is based on the notion that the primary decisions the individuals make are when to change the purchases and sales of commodities and bonds. These decisions result in money holdings that are held passively. By contrast, the "active" decision is not when to buy or sell "money."

The bond market in the IS–LM diagram

This section uses Walras's law to derive the locus of points in the IS–LM diagram at which there will be equilibrium in the bond market. Figure 19.1 shows the standard IS and LM curves and assumes, for simplification, a closed economy with continuous equilibrium in the labor market with full employment. Given this simplifying assumption, the equilibrium

[10] The money market is an "image" provided by the merged reflection of two (or several) entities or figures (markets) standing in front of a mirror, with Walras's law acting as the mirror. Looking at the composite reflection only provides a great deal of information – but not necessarily on the separate elements – about each of the figures (markets) facing the mirror, but is itself not independent of the existence and nature of the mirror, or of the figures (markets).

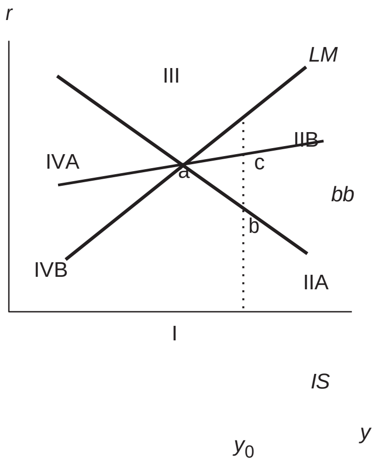


Figure 19.1

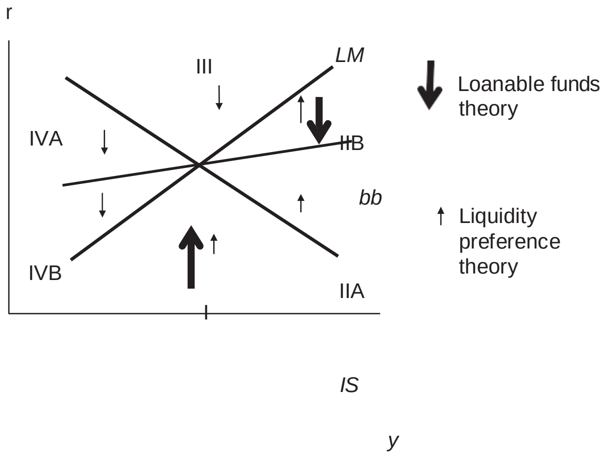


Figure 19.2

curve for the labor market has been omitted from this figure. This figure also assumes an exogenously given expected inflation rate set at zero, so that r ~~*R*~~.

The bb curve in Figure 19.1 specifies the combinations of (r, y) which maintain equilibrium in the bond market, so that E_b is zero at all points along it. By Walras's law, the equilibrium in the commodity and money markets shown by the intersection of the IS and LM curves at point a ensures that the bond market will also be in equilibrium at a. Therefore, the bb curve must pass through the intersection of the IS and LM curves, so that all three curves must pass through the common point a.

The IS–LM Figures 19.1 and 19.2 are divided into four quadrants. In quadrant I, there exists excess demand for commodities since, for a given income, interest rates are lower than those specified by the IS curve so that investment is higher than required for equilibrium. That is, $E_c > 0$. There is also excess demand for money, since the interest rate is lower than specified by the LM curve, so that the speculative demand for money is too high compared with that in equilibrium. That is, $E_m > 0$. By Walras's law for our assumed economy, with $E_c > 0$ and $E_m > 0$, we must have $E_b < 0$. That is, there will exist an excess supply of bonds at all points in this quadrant. Therefore, the bb curve (on which, by definition, E_b 0 at every point) cannot pass through quadrant I.

By similar reasoning, it can be shown that in quadrant III, $E_c < 0$ and $E_m < 0$, so that, by Walras's law, we must have $E_b > 0$. Hence, the bb curve also cannot pass through quadrant III.

To summarize, the excess demand functions for quadrants I and III are:

I: $E_c > 0$, $E_m > 0$, $E_b < 0$ III:
$E_c < 0$, $E_m < 0$, $E_b > 0$

Since the bb curve with $E_b = 0$ along it cannot pass through quadrants I and III, it must pass through quadrants II and IV. The latter are further divided into regions IIA and IIB, and IVA and IVB, by the depicted position of the bb curve. To illustrate what happens in regions IIA and IIB, we first examine the point b in Figure 19.1. At this point, the interest rate is too low for equilibrium in the bond market, so that the existing bond prices are too high and there is inadequate demand (excess supply) for bonds at this lower-than-equilibrium interest rate. At point c in the region IIB, the interest rate is too high and bond prices too low for equilibrium in the bond market, so that there would be too much demand (positive excess demand) for bonds at the higher-than-equilibrium interest rate. Similar reasoning can be used to separate regions IVA and IVB. The excess demand functions for regions II and IV are:

IIA: $E_c < 0$, $E_m > 0$ and $E_b < 0$

IIB: $E_c < 0$, $E_m > 0$ but $E_b > 0$

IVA: $E_c > 0$, $E_m < 0$ and $E_b > 0$

IVB: $E_c > 0$, $E_m < 0$ but $E_b < 0$.

Since IIA has $E_b < 0$ while IIB has $E_b > 0$, the separating region between them must have $E_b = 0$. This is the requirement for points on the bb curve, so that the bb curve does differentiate between two distinctive parts of region II. Similarly, IVA and IVB are separated by the locus of points at which $E_b = 0$. Hence, the bb curve passes through regions II and IV and not through I and III. This argument does not establish whether the bb curve will have a positive or a negative slope in the (r, y) space, though we have shown a positively sloping curve. The nature and magnitude of the slope will depend on the coefficients of the IS and LM equations, as can be seen from (21).

We can now examine the impact on the bb curve of changes in the exogenous variables and parameters of the model. We do so explicitly only for the policy variables of the fiscal deficit and the money supply. Start with an increase in the fiscal deficit in Figure 19.1. This would shift the IS curve to the right (not shown). With the LM curve taken to be independent of the fiscal deficit, by Walras's law, the bb curve must shift to pass through the intersection of the new IS and the initial LM curves. Hence, the bb curve will also shift to the right. The intuitive reason has to do with the bond financing of the deficit, which increases the supply of government bonds in the economy. Equilibrium in the bond market requires either higher income or higher interest rates to generate a corresponding increase in the demand for bonds.

But suppose, instead, that the money supply had increased. This would shift the LM curve to the right. Walras's law again implies that the bb curve must shift to pass through the new intersection of the initial IS curve and the new LM one. Hence, the bb curve shifts downwards. The intuitive reason for this is that the increased money supply is traded by the public for bonds, thereby increasing the demand for bonds. This raises bond prices and lowers the interest rate.

The dependence of the bb curve on shifts in both the IS and the LM curves is also apparent from equation (21), which shows that the excess demand for bonds depends on the fiscal deficit and the money supply. The bond market is thus the implicit link between the IS and the LM curves in the IS–LM model.

We derived the bb curve under the assumption of full employment and therefore of constant real output. If the labor market were in disequilibrium, Walras's law would require a modification of our arguments, without, however, a change in the nature of the bb curve. Further, in general equilibrium, the labor market would clear, so that the bb curve must still pass through the intersection of the IS and LM curves.

19.6.1 Diagrammatic analysis of dynamic changes in the rate of interest

The loanable funds theory asserts that the dynamic movement of the interest rate is determined by the excess demand for bonds, while the liquidity preference theory asserts that it is determined by the excess demand for money. Figure 19.2 uses arrows to show the movements in the interest rate implied by the liquidity preference theory, which asserts that $\partial R/\partial(E_m{}^{\mathrm{d}}) > 0$, and the loanable funds theory, which asserts that $\partial R/\partial(E_b{}^{\mathrm{d}}) < 0$. The movements implied by the former theory are shown by light arrows indicating the direction of movement, while those implied by the latter are shown by the heavy arrows. Both theories predict an increase in the interest rate in quadrant I, and both theories predict a decline in it in quadrant III. Further, in region IIA, both theories indicate a rise in the interest rate, and in region IVA both theories indicate a fall in it. However, note that the implied magnitude of the change in the interest rate could differ between the theories.

The especially interesting regions are IIB and IVB. Region IIB has $E_m > 0$ and $E_b > 0$. Therefore, the liquidity preference theory predicts a rise in the interest rates while the loanable funds theory predicts a fall. In region IVB, $E_m < 0$ and $E_b < 0$, so that the liquidity preference theory predicts a fall in the interest rate and the loanable funds theory predicts a rise. Consequently, the dispute between these theories is not trivial even in terms of the sign of the movements in the interest rates in the economy.

The dispute is of importance even in quadrants I and III and regions IIA and IVA, where the two theories predict a similar direction of change in the interest rate, but the dynamic speed of movements in the rate will be sensitive to the actual relationship and is likely to differ. Hence, a choice has to be exercised between the two theories for both qualitative and quantitative dynamic analyses, though not for the general equilibrium case.

Classical heritage: the loanable funds theory of the rate of interest

The traditional classical economists (prior to Keynes) had generally favored the specification of the overall equilibrium in terms of the bond, money and labor markets, with the labor market determining employment and, through the production function, output; the bond market determining the interest rate, and the market for money determining the price level. Its theory of the determination of the interest rate (or its inverse, the price of bonds for bonds specified as consols which have fixed coupon payments payable perpetually) was known as the *loanable funds theory*. It asserted that the interest rate was determined in the bond market by equilibrium between the demand and supply of "loanable funds," which was its synonym for the current term "bonds." Given the discussion so far in this chapter, we can distinguish the following three aspects of the loanable funds theory:

1 Partial equilibrium (short-run) determination of the interest rate.
2 General equilibrium (long-run) determination of the interest rate.
3 Dynamic movement of the interest rate.

The traditional classical economists did not have a distinction (until the advent of Fisher's equation) between the real and nominal interest rate, so that they normally referred to the market interest rate R as the determinant of investment and saving. Following their pattern of analysis, we will specify the investment function in the following as $i(R)$, rather than our usual $i(r)$.[11] They also did not have a government sector with outstanding bond-financed budget deficits.[12] Further, the role of financial intermediaries and the monetary base in the money creation process were not fully understood.[13]

Given these simplifications, the demand and supply functions of the loanable funds theory were:

$$Pf^{s} = Ps(R, y) + [M^{s} - M^{d}]$$

$$Pf^{d} = Pi(R)$$

Therefore, *the short-run* (i.e. partial equilibrium) *determination of the interest rate* according to the loanable funds theory was specified by:

$$s(R, y) + (1/P)[M^{s} - M^{d}] = i(R) \qquad (23)$$

so that:

$$R = \phi(P, y;(M - M^{d})) \qquad (24)$$

which allows both the commodity market *and the money market* shifts to change the interest rate.

The *long-run version of the loanable funds theory* assumed general equilibrium in the economy. Therefore, for this version, $M—M^{d} =0$, so that the long-run loanable funds theory became:

$$i(R) = s(R, y) \qquad (25)$$

[11] This would also be correct under the Fisher equation if the expected inflation rate was zero.

[12] The loanable funds theory was formulated in a period when the size of the government was relatively small. In any case, the formal bond market was not sufficiently developed for governments to sell bonds in them. Often, any borrowing by the government was through private arrangements with individual banks and private financiers.

[13] The traditional classical theories were also formulated for an era in which the formal financial sector was

relatively insignificant. Until the Second World War, the market for credit was dominated by firms ("ultimate borrowers") raising funds for investment and savers ("ultimate lenders") lending out of savings, so that it was common for the role of financial intermediaries to be ignored or, in any case, not properly integrated into the theory of the rate of interest.

Further, long-run equilibrium in the commodity and labor markets ensures that output will be at the full-employment level (y^f), so that (25) becomes:

$$i(R) = s(R, y^f) \qquad (26)$$

That is, the long-run interest rate is determined by the equality of investment and saving at full-employment output, so that its main determinants are the propensity to save, the production capacity of the economy, and investment. In particular, this interest rate is not altered by shifts in the demand or supply of money.

For the *dynamic movement of the interest rate* when there is disequilibrium in the economy, the loanable funds theory asserted that the interest rate is determined by the excess demand or supply of loanable funds: it falls (while the bond price rises) if there is an excess demand for loanable funds, and rises (while the bond price falls) if there is an excess supply of loanable funds. Therefore, this theory's assertion for dynamic adjustments in the interest rate is:

$$R = f(E_b{}^d) \qquad \partial R/\partial(E_b{}^d) < 0 \qquad (27)$$

Note that, since changes in the money supply and demand alter the excess demand for bonds, they will also affect the dynamic path of the interest rate.

To conclude on the relevance of the excess money supply in changing the nominal interest rate, only the long-run version of the loanable funds theory asserted its irrelevance for the determination of the interest rate. However, this long-run version, which is the statement that the interest rate is determined by saving at full employment and investment, is the one usually remembered as the statement of the loanable funds theory.

Adapting the loanable funds theory to the modern economy requires the introduction of the government sector, the central bank and the financial sector into the component functions of this theory, as presented earlier in this chapter.

Loanable funds theory in the modern classical approach

In recent years, the modern version of the classical paradigm has reasserted continuous market clearance for the labor markets, as for the other markets, and with its assumption of rational expectations has further asserted the possibility of disequilibrium (due to expectational errors) in any market as at best a very transitory state. That is, with the labor and money markets clearing continuously, there would exist full employment in the economy and the excess demand in the money market would be zero. Consequently, for the modern classical school, the theory of interest reverts to the long-run version of the traditional loanable funds theory, with the difference that it is now intended to be not only the long-run theory but also the short-run theory of the rate of interest as far as systematic or anticipated changes in the money supply are concerned.[14] However, such a short- run theory could still diverge from its long-run version because of random influences operating on the economy in the short run. These cannot be anticipated under rational expectations and would cause a divergence of the short-run interest rate from its long-run level.

[14] The latter is a departure from traditional classical economics, as a comparison of the doctrines of the modern classical school with the quotation from Hume in the next subsection clearly shows.

The modern classical version of the loanable funds theory, therefore, extends and differs from the traditional classical one in various ways. Among the differences are:

1 The role of financial intermediaries, as discussed earlier.

2 The addition of Fisher's equation connecting the real and nominal interest rates in perfect capital markets.

3 The distinction in the modern version between the anticipated and unanticipated values of the relevant variables, among which are the money supply, the other determinants of aggregate demand and the rate of inflation. Anticipated money supply increases cause anticipated inflation without changing the real interest rate and, therefore, increase the nominal rate by the anticipated rate of inflation, as specified by the Fisher equation. Unanticipated money supply growth lowers the real rate and will lower the market rate of interest.

4 Ricardian equivalence, which makes national saving independent of the (anticipated) fiscal deficit and therefore removes such deficits from the determinants of the demand and supply of loanable funds. In this case, anticipated deficits would not affect the interest rate.

5 In the short run, the traditional classical economists allowed deviations from full employment under the impact of money supply changes and the impact of these changes on saving. The modern classical economists allow such a deviation for only unanticipated money supply changes. Therefore, the short-run deviations of output from its full employment level under the impact of anticipated money supply changes could, in the short run, affect the interest rate under the traditional version of the loanable funds theory but not under its modern version.

Note that outside the confines of long-run general equilibrium analysis, the interest rate is not merely the reward for postponing consumption, it is also the return on lending, which is the act of parting with liquidity, i.e. not holding money. The latter was a major contention of Keynes and is a fundamental part of Keynesian beliefs.

David Hume on the rate of interest

David Hume occupies a special place in the macroeconomic theory of the rate of interest because he specified the main elements of the preceding traditional classical theory at an early stage in its development. He expressed some of the basic elements of the traditional classical short- and long-run theories in his essay *On Interest*, published in 1752. He stated its long-run version as:

> For, suppose that, by miracle, every man in Great Britain should have five pounds slipped into his pocket in one night; this would much more than double the whole money that is at present in the kingdom; yet there would not next day, nor for some time, be any more lenders, nor any variation in the interest. And were there nothing but landlords and peasants in the state, this money, however abundant, could never gather into sums, and would only serve to increase the prices of everything, without any further consequence. The prodigal landlord dissipates it as fast as he receives it; and the beggarly peasant has no means, nor view, nor ambition of obtaining above a bare livelihood. The overplus of borrowers above that of lenders continuing still the same, there will follow no reduction of interest. That depends upon another

> principle; and must proceed from an increase of industry and frugality of arts and commerce.
>
> The greater or less quantity of it [money] in a state has no influence on the interest. But it is evident that the greater or less stock of labor and commodities must have a great influence; since we really and in effect borrow these, when we take money upon interest.
>
> (Hume, *Of Interest*, 1752).

Hence, for the long run, Hume asserts that changes in the money supply have no impact on the (long-run) interest rate, which is determined by the real factors of labor supply, productivity of investment, and saving. For the short run, they also have no impact if increases in the money supply are spent wholly on commodities but not saved. Hume next considers the *dynamics* of a change in the money supply in the following statements.

> Another reason of this popular mistake with regard to the cause of low interest, seems to be the instance of some nations, where, after a sudden acquisition of money, or of the precious metals by means of foreign conquest, the interest has fallen. … it is natural to imagine that this new acquisition of money will fall into a few hands, and be gathered into large sums, which seek a secure revenue, either by the purchase of land or by interest. … The increase of lenders above the borrowers sinks the interest, and so much the faster if those who have acquired those large sums find no industry or commerce in the state, and no method of employing their money but by lending it at interest. But after this new mass of gold and silver has been digested, and has circulated through the whole state, affairs will soon return to their former situation. … The whole money may still be in the state, and make itself felt by the increase of prices; but not being now collected into any large masses or stocks, the disproportion between the borrowers and lenders is the same as formerly, and consequently the high interest returns.
>
> (Hume, *Of Interest*, 1752).

The salient points of Hume's conclusions can be summarized in modern terminology as follows.

1 The long-run real equilibrium interest rate is determined by saving at full-employment output and productivity of investment.

2 The long-run interest rate is invariant to changes in the money supply. A long-run decline of the rate of interest is brought about by a decline in the productivity of investment and not by increases in the money supply.

3 However, in the short run, the real interest rate is lowered (and output increased) by those increases in the money supply that increase the supply of loanable funds in the economy, as occurs in the indirect transmission mechanism, but not by those that do not, as occurs in the direct transmission mechanism. The structure of the economy and the mode of introducing additional money balances into the economy determine which mechanism is the relevant one, and how long the reduction of the interest rate will last.

Note that the only thing that seems to be missing from Hume's arguments is the distinction between the real and the nominal interest rates: Hume did not have Irving Fisher's equation, proposed early in the twentieth century, relating the nominal interest rate to the expected rate of inflation.

Keynesian heritage: the liquidity preference theory of the interest rate

Keynes's *General Theory* (1936) challenged the loanable funds theory on the grounds that the interest rate was not the reward for saving but was rather an inducement to part with liquidity. He summarized his views in the statement:

> [Once the individual has made his decision on consumption versus saving out of his income], there is a further decision which awaits him, namely, in what form he will hold the command over future consumption which he has reserved, whether out of his current income or from previous savings. Does he want to hold it in the form of immediate, liquid command (i.e. in money or its equivalent)? Or is he prepared to part with immediate command for a specified or indefinite period. …
>
> It should be obvious that the rate of interest cannot be a return to saving or waiting as such. For if a man hoards his savings in cash, he earns no interest, though he saves just as much as before. On the contrary,…, the rate of interest is the reward for parting with liquidity for a specified period. …
>
> Thus the rate of interest at any time, being the reward for parting with liquidity, is a measure of the unwillingness of those who possess money to part with their liquid control over it. The rate of interest is not the "price" which brings into equilibrium the demand for resources to invest with the readiness to abstain from present consumption. It is the "price" which equilibrates the desire to hold wealth in the form of cash with the available quantity of cash. … If this explanation is correct, the quantity of money is the other factor, which, in conjunction with liquidity preference, determines the actual rate of interest in given circumstances.
>
> (Keynes, 1936, pp. 166–8).

First, consider Keynes's argument in terms of its general notion that the interest rate is the reward for parting with liquidity. This is definitely true in a world with uncertainty. Savers have a choice as to the form in which to hold their savings. They may hold these in a monetary form or lend it. If the level of the interest rate determines their division of savings into money balances versus loans, the interest rate can be called the reward for parting with liquidity in the process of lending. However, if the interest rate also influences the level of savings, then it may also be called a reward for postponing consumption. Both cases apply in the real world.[15]

Now consider Keynes's argument formally in terms of the equilibrium relationship of the monetary sector. As shown in Chapter 2 above, Keynes's money market equilibrium relationship for an exogenously given money supply M is:

$$M = kPy + L(R) \qquad (28)$$

Equation (28) determines R *if* it is assumed that P and y are exogenously given. This is not true of the Keynesian model and is not true for Keynes's own ideas in general. In his theory, output, interest and prices were determined simultaneously so that R is not determined merely by (28): it is also influenced by the saving and investment decisions of the expenditure sector

[15] However, several empirical studies show the impact of interest rates on saving to be insignificant or of little importance.

as well as by the labor market structure. Hence, the interest rate is not merely the reward for parting with liquidity, even though that may seem to be the most proximate or closely related cause.

Dynamics of the liquidity preference theory

According to Keynes's liquidity preference theory, the dynamic movements of the interest rate are determined by the excess demand for money. Hence, it was asserted that:

$$R = f(E_m{}^{\mathrm{d}}) \qquad \partial R/\partial(E_m{}^{\mathrm{d}}) > 0 \tag{29}$$

so that:

$$\partial R/\partial(M^{\mathrm{d}} - M^{\mathrm{s}}) > 0 \tag{30}$$

The argument behind this assertion runs as follows. According to Keynes an excess demand for money by individuals would make them sell bonds in order to obtain the extra money balances they want. These bond sales will lower bond prices, which will raise the interest rate.

Comparing the liquidity preference and the loanable funds theories of interest

A lengthy controversy flared up in the 1950s and early 1960s as to whether the traditional classical loanable funds and the Keynesian liquidity preferences theories were identical or different. Given our analysis so far, this question can be answered for the separate categories of general equilibrium and dynamic analyses.

General equilibrium analysis

Our earlier analysis implies that, given Walras's law, it is immaterial whether the *general equilibrium solution* of the macroeconomic model is obtained from (i) the traditional classical set consisting of the money, bond and labor markets, or (ii) the Keynesian set consisting of the commodities, money and labor markets. Each set would give the *same* general equilibrium values of all the endogenous variables, even though the two sets, *prima facie*, would seem to be quite different.

Note that the current practice in macroeconomic modeling is to specify the complete model in terms of the commodity, money and labor markets. The selection of the bond market for omission is partly due to the tradition set by Keynes in *The General Theory*, reinforced by Hicks in his interpretation of Keynes in the form of the IS–LM model, and partly because most countries do not publish adequate and reliable data on the amounts of bonds in the aggregate and for many types of bonds (and loans) in the economy. By comparison, the data on output, money and employment – and their related variables – is usually made available in great detail and with more or less of an attempt at consistency over time.

Dynamic analysis

For the dynamics of interest rate movements, the selection of the sector in which the rate of interest is determined is highly relevant. This has already been shown above in various places. The following pulls it together in one place.

The dynamic version of the liquidity preference theory is that the changes in the rate of interest are determined by the excess demand for money, with a positive relationship. That is:

$$\mathrm{d}R/\mathrm{d}t = W(E_{mt}) \qquad W^{J} > 0$$

$$= W(M^{\mathrm{d}}_{t} - M^{\mathrm{s}}_{t}) \tag{31}$$

The dynamic version of the loanable funds theory is that changes in the rate of interest are determined by the excess demand for bonds, with an inverse relationship:

$$\mathrm{d}R/\mathrm{d}t = \varphi(E_{bt}) \qquad \varphi^{J} < 0$$

$$= \varphi(P\{i(R) - s(R)\} - \{M^{\mathrm{d}}{}_{t} - M^{\mathrm{s}}{}_{t}\}) \tag{32}$$

Equations (31) and (32) generate different time paths for the interest rate. To illustrate an extreme scenario, if the money market is in equilibrium but the bond market is not, the liquidity preference theory (31) would have E_{mt} 0 and imply that the interest rates will not change, but the loanable funds theory (32) wōuld have a non-zero excess demand for commodities and bonds and would, therefore, imply changes in the interest rate.

Neutrality versus non-neutrality of the money supply for the real rate of interest

Neutrality of money under an exogenous money supply

The real rate of interest is a real variable. As such, the modern classical analysis (Chapter 14) implies that in the long-run general equilibrium (without errors in price or inflationary expectations) the real interest rate will be invariant to changes in the (nominal) money supply, since such changes in the money supply produce proportionate changes in the price level without changing the real money supply. Hence, changes in the money supply do not change the long-run real interest rate in the modern classical analysis.

However, for the short run, the modern classical school allows errors in expectations to affect the real interest rate. This effect can be illustrated by the adaptation of the Friedman–Lucas output supply function to interest rate determination, as in:

$$r_t = r^{*} + \gamma(M_t - EM_t) \qquad \gamma < 0 \tag{33}$$

where r^* is the long-run value of the real rate of interest. Hence, in the modern classical models, unanticipated increases in the money supply are not neutral in the short run; they lower the real interest rate. But anticipated increases in the money supply do not change the real interest rate and are neutral.

The new Keynesian models incorporate several elements, such as imperfect information and sticky prices, which lead to the non-neutrality of money with respect to output and employment. They also do so for the real interest rate.

Neutrality of monetary policy under a Taylor interest rate rule

Chapter 13 presented the general form of the Taylor interest rate rule as:

$$r^{\mathrm{T}}{}_{t} = r^{\mathrm{LR}} + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \qquad \alpha,\ \beta > 0 \tag{34}$$

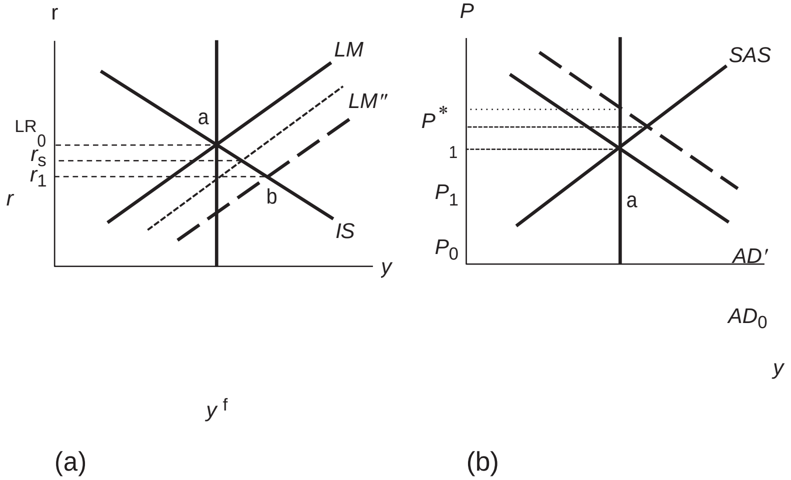


Figure 19.3

where r^T is the real interest rate target, y is real output, y^f is full-employment output, π is the actual inflation rate, π^T is the inflation rate desired by the central bank, and the subscript t refers to period t. π^T is called the *target inflation rate*. As shown in Chapter 14, in the long run, $y_t = y^f$ and $\pi_t = \pi^T$, so that $r = r^{LR}$. Hence, the long-run interest rate will be invariant with respect to monetary policy even if the central bank sets interest rates. However, in the short term, the real rate in the financial markets will depend on the interest rate set by the central bank, so that monetary policy will not be neutral in terms of its effect on the real interest rate on bonds.

Diagrammatic analysis of the role of the money supply in the determination of the rate of interest

Figure 19.3a shows the long-run equilibrium interest rate at $r^{LR}{}_0$. In the long run, an increase in the money supply will shift the LM curve out to LM^J, increasing demand to the point b, thereby causing a long-run price increase (from P_0 to $P^*{}_1$ in Figure 19.3b) which is sufficient to return the LM curve in Figure 19.3a back from LM^J to LM and the general equilibrium real interest rate back to $r^{LR}{}_0$. The long-run equilibrium real interest rates and output are therefore invariant to the nominal money supply increase.

But, in the short run, assuming that the money market adjusts instantly while the commodity market is slower to adjust, the economy would initially lower the real interest rate to $r_{1.}$ If the economy proceeds along a short-run supply curve SAS – either for Keynesian or for modern classical reasons (with errors in relative price expectations) – the monetary expansion will shift the demand curve to AD^J in Figure 19.3b, and lead to a price increase from P_0 to P_1 (rather than to $P^*{}_1$). This will mean, in Figure 19.3a, a shift in the LM curve up from LM^J to only LM^J (rather than back to LM), and yield a short-run rate of interest r_s. Therefore, increases in the money supply can have both immediate (from $r^{LR}{}_0$ to r_1) and short-run effects (from r_s to $r^{LR}{}_0$) on the interest rate in the economy, but not in the long run.

Determinants of the long-run ("natural") real rate of interest and the non-neutrality of fiscal policy

We now look at the factors that can change the long-run real rate of interest. Figure 19.4 shows the determination of the long-run real interest rate r^{LR}. This is given by the intersection of the

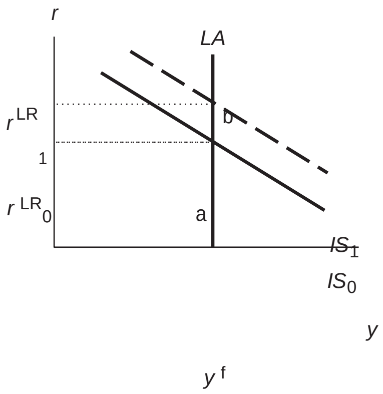


Figure 19.4

IS curve and the long-run aggregate supply curve LAS. Shifts in either of these will change the long-run rate of interest in the economy. Therefore, productivity shifts in the economy as well as its saving and investment behavior will shift the long-run real rate of interest. So will government expenditures and tax rates in the standard IS–LM model, since a deficit in this model shifts the IS curve to the right. This effect is shown in Figure 19.4 with a rightward shift in the IS curve from IS_0 to IS_1 due to a budget deficit, with a consequent increase in the long-run interest rate from $r^{LR}{}_0$ to $r^{LR}{}_1$. That is, higher deficits produce higher real rates of interest, and eliminating them will lower the real interest rate in the long run. The LM curve is irrelevant to the determination of the long-run real rate of interest, so that it has not been drawn in Figure 19.4. We conclude from this figure that, for the closed economy and a given production function, the long-run real interest rate is determined by the equality of investment and national saving, which equals private saving less the fiscal deficit. A decrease in national saving reduces the loanable funds available for investment and raises the real interest rate.

However, the addition of Ricardian equivalence to the IS–LM framework implies, as shown in Chapters 13 and 14, that private saving increases by the amount of the deficit, so that national saving does not change as a result of a deficit. Therefore, budget deficits will not shift the IS curve, so that, in Figure 19.4, the long-run real interest rate will remain at $r^{LR}{}_0$ irrespective of the deficit. Hence, given Ricardian equivalence, the long-run real interest rate will be invariant to fiscal policies. A cautionary note about this conclusion is needed: if Ricardian equivalence does not hold (see Chapter 14 on its doubtful empirical validity), fiscal deficits do alter the long-run real interest rate.

Other determinants of the long-run real rate of interest

The other determinants of the long-run real interest rate include the following.

1 The efficiency of financial intermediation and innovations in it, since these affect the efficiency with which savings are collected and allocated among investment projects.

2 Innovations and the rate of technical change generally in the economy, which affect the growth of output.

3 The openness of the economy to world markets and international capital flows, since such flows can cover a saving–investment gap.

Empirical evidence: testing the Fisher equation

For tests of the Fisher equation of the nominal interest rate, note that an unanticipated increase in the money supply lowers the short-run real interest rate as the economy moves down the SAS curve, thereby tending to lower the nominal rate, while the resulting inflation, through anticipations, causes an increase in the nominal rate. Further, the Mundell–Tobin effect, discussed earlier in this chapter, implies a negative impact of the expected inflation rate on the real interest rate since the former is the opportunity cost of holding money and reduces the demand for real balances.

A test of the Fisher equation is provided by Crowder and Hoffman (1996). Since an increase in the interest rate due to inflation increases the tax payments on interest receipts, Crowder and Hoffman argued that the empirical estimates of the Fisher equation should have an estimated coefficient of the rate of inflation between about 1.3 and 1.5, rather than unity, as implied by the tax-free Fisher equation, even though many empirical studies had found this coefficient to be less than unity. Since the data on nominal interest rates and inflation tends to be non-stationary, Crowder and Hoffman use the Johansen cointegration technique. Their estimating equation is based on a generalized form of the Fisher equation derived from intertemporal utility maximization subject to a budget constraint, and is stated as:

$$R_t(1-\tau_t) = r_t + E_tOp_{t+1} + 0.5\,\mathrm{var}_tOp_{t+1} - \gamma\ \mathrm{cov}_t(Oc_{t+1},\ldots, Op_{t+1}) \tag{35}$$

where p is the log of the price level, c is the log of consumption, E_t is the expectations operator conditional on information in t, R is the nominal rate and r the real one, τ is the tax rate and γ is the coefficient of relative risk aversion. In this study, the estimated coefficients for the expected rate of inflation were in the range from 1.34 to 1.37 and, when adjusted for the tax rates, in the range 0.97 to 1.01, so that this study supported the Fisher effect.

The Fisher effect incorporated into the theory of the term structure of interest rates allows the estimation of the expected rate of inflation from the yield curve. This derivation will be discussed in the next chapter.

Testing the liquidity preference and loanable funds theories

Finding an empirical test that would truly distinguish between the liquidity preference and the loanable funds theories poses quite a problem. The strong distinction between these theories only emerges in the hypothetical/analytical long-run general equilibrium state of the economy, which is extremely difficult to test for since the available data never relates to this state of the economy. The tests therefore have to be of the theories' short-run versions. But the short-run versions of both theories imply that excess money demand does affect the interest rate. To illustrate the nature of the problem, the loanable funds theory (LF) asserts that:

$$\text{LF:}\quad R = f(E_b) \tag{36}$$

which, by Walras's law with labor market clearance, becomes:

$$\text{LF:}\quad R = f(-E_c - E_m) \tag{37}$$

The liquidity preference theory (LP) asserts that:

$$\text{LP:}\quad R = f(E_m) \tag{38}$$

Hence, E_m occurs as a determinant of the interest rate in both theories.

Alternatively, using Walras's law to replace E_m by $(-E_c - E_b)$, we have:

$$\text{LF:} \quad R = f(E_b) \tag{39}$$

$$\text{LP:} \quad R = f(-E_c - E_b) \tag{40}$$

In this case, E_b occurs as a determinant of R in both theories, even the liquidity preference theory.

Therefore, the problem in estimations meant to distinguish between the loanable funds and liquidity preference theories arises because both E_b and E_m affect the interest rate in both theories.

Empirical findings

Feldstein and Eckstein (1970) sought to provide an application of the liquidity preference theory, combining it with the Fisher equation for the relationship between the nominal interest rate and the expected inflation rate. The liquidity preference theory implies that the rate of interest depends in the short run upon the excess demand for money. Since the demand for money depends upon income, the excess demand for money was represented through the use of the monetary base and national income among the explanatory variables, with the former capturing the effect of an increasing money supply and having a negative expected coefficient, while the latter captures the effect of increasing money demand and has a positive expected coefficient. The expected rate of inflation was proxied by a distributed lag autoregressive model. Among the results reported by these authors are:

$$R_t = -11.27 - 6.76 \ \ln \ \text{M0}_t + 6.03 \ \ln \ y_t + 0.275\pi_t + \Sigma_j \alpha_j \pi_{t-j} \tag{41}$$

where $j = 1, 2, \ldots, 23$, $\Sigma_j \alpha_j = 3.41$, with $\alpha_1 = 0.289$ and $\alpha_{23} = 0.020$, $R^2 = 0.982$.

In (41), R was a corporate bond rate (the yield on seasoned Moody's Aaa industrial bonds), M0 was the real per capita monetary base and y was real private GDP per capita. All the coefficients in (41) were significant and the mean lag for the impact of inflation on the interest rate was 8.14 quarters. These results are consistent with the liquidity preference theory.

Feldstein and Eckstein extended (41) to include privately held Federal government debt, and reported the estimates as:

$$R_t = -16.68 - 9.08 \ \ln \ \text{M0}_t + 8.24 \ \ln \ y_t + 2.78D_t + 0.27\pi_t + \Sigma_j \alpha_j \pi_{t-j} \tag{42}$$

where $\Sigma_j \alpha_j = 3.93$, $j = 1, 2, \ldots, 23$, $R^2 = 0.985$, the mean lag for the inflation impact

was

7.90 quarters and D was real per capita privately owned Federal government debt. In (42), the coefficient of government debt is positive since, as explained by the loanable funds theory, an increase in the supply of bonds, i.e. the demand for loanable funds, will raise the interest rate. In both (41) and (42) the mean lag for the impact of inflation on the interest rate is about 8 quarters and therefore quite long. Further, while the increase in the monetary base directly lowers the nominal interest rate, its indirect impact on inflation raises this rate. The reported coefficients imply that there does not exist short-term neutrality of money with respect to the real rate, at least not for periods up to 8 quarters.

While (41) represents the liquidity preference theory, (42) combines elements of both the liquidity preference and the loanable funds theories, the latter through its inclusion of government debt. Since the coefficients of both the monetary base and government bonds are significant, the above estimates support the general version of the determination of the interest rate as given by a broader commodities–money–bonds model with Walras's law, rather than providing a rejection of either of the rival theories.

The general conclusion of the Feldstein and Eckstein (1970) study was that the rise in the interest rate between 1954 and 1965 was due more to decreasing liquidity than to inflation, but that inflation was more important from 1965 to 1969. The relatively slow growth of the public debt through the period held back the increase in the interest rate. Further, the direct impact of the changes in the monetary base and the government debt on the interest rate occurred within one quarter, so that there was not a significant lag in these effects, while the impact of inflation took place over 23 quarters, with a mean lag of about eight quarters. It is not clear whether such a long lag arises from a lag in the adjustment of the nominal interest rate to the expected rate of inflation, or from a lag in the expected rate in adjusting to the actual inflation rate. However, what is clear from this study is that Fisher's relationship between the interest rate and expected inflation must not be omitted in estimations of the nominal rate, even in studies based on the liquidity preference approach.

Sargent (1969), and Echols and Elliot (1976),[16] provided applications of the loanable funds theory. Sargent (1969) started with the identity:

$$R \equiv r* + (r - r*) + (R - r) \qquad (43)^{17}$$

where R is the nominal rate of interest (holding period yield on a bond). r is the real rate (the nominal rate less the expected rate of inflation over the holding period), and was called by Sargent the market real rate of interest. r^* is the rate of interest that equates investment and saving and corresponds to the "normal rate of interest" in Wicksell, as explained in Chapter 2. As in Wicksell's analysis, it was made a function of the excess of investment over saving. Its use by Sargent was meant to capture the loanable funds theory. $(r\ r^*)$ is the deviation of the market real rate from the normal rate. As in Wicksell's analysis, this deviation depended upon the excess supply of money created through the bank's operations, which increases the supply of loans over that through saving. From the Fisher equation, $(R\quad r)$ equaled the expected rate of inflation in commodity prices. Sargent represented expectations by a distributed lag model.

The normal rate r^* was specified as a function of the excess demand for loans, which was defined as desired real investment less desired real saving. Investment was specified as a function of this rate and OX, while saving was a function of this rate and X, where X is real output. Among the reported estimates[18] are the following ones for the *ten-year bond yield*:

$$R_t = 7.1338 + 0.0099OX_t - 0.0456X_t - 2.0151(Om^*{}_t/m^*{}_{t-1})$$

$$+ 3.8764\ \Sigma_i\ 0.97^{i-1}(Op_{t-i}/p_{t-i-1}) - 1.9849(0.97)^t \qquad (44)$$

Adjusted $R^2 = 0.9298, i = 1, 2,\ldots, t - 1.$

[16] The approach and results of this study will be examined in Chapter 20 in the context of the term structure of interest rates.

[17] The symbols in this equation have been altered for consistency with our symbols in this chapter.

[18] The Hildreth–Lu procedure was used to correct for serial correlation.

where R was the ten-year nominal bond yield, X was real GDP, m^* was real money supply and p was the commodity price index. All the coefficients were significant, except for OX_t, and the signs were consistent with the assumed hypotheses. An increase in the real money supply reduced the real rate of interest, with a 10 percent increase in the real money supply reducing the nominal interest rate by 20 basis points.

The estimated results for the *one-year bond yield* were:

$$R_t = 1.4396 + 0.0182OX_t - 0.0405X_t - 6.0260(Om^*{}_t/m^*{}_{t-1})$$

$$+ 6.4716\ \Sigma_i\ (0.98)^{i-1}(Op_{t-i}/p_{t-i-1}) + 4.4933(0.98)^t \qquad (45)$$

Adjusted R^2 0.9298, i 1, 2,..., t 1.
= = −

In (45), the output level, changes in the money supply and the expectations variable were all significant and had the expected signs. The coefficient of the rate of change of output was insignificant in both (44) and (45). This variable was meant to capture the inclusion of the demand for loanable funds through investment.

Equations (44) and (45) show that both the money supply and the inflation rate have greater impact on the one-year rate than on the ten-year rate. Both indicate very long lags in the impact of inflation on the interest rate, possibly because of long lags in the process of expectations formation, though this result may have been due to the representation of expectations by a distributed lag function rather than by rational expectations.

Both equations include changes in the money supply, which is an element in the liquidity preference theory. Further, the inclusion of output could capture elements of money demand determination. Hence, it cannot be claimed that these equations exclude elements of liquidity preference. However, they also include elements of the traditional classical loanable funds theory. We tend to view them, as we did (41) and (42), as being consistent with the general macroeconomic model with Walras's law, and therefore with the interest rate moving in response to disequilibrium in both the money and bond markets.

Conclusions

This chapter has focused on the underlying interest rate in the economy, while leaving the study of the term and risk structure of interest rates to the next chapter. There are two main theories of this underlying rate. The loanable funds theory is associated with traditional classical economics and asserts that the interest rate is determined in the market for loanable funds – designated as bonds or credit in modern macroeconomic models. The liquidity preference theory is associated with Keynes and Keynesian economics, and asserts that the interest rate is determined by the equilibrium in the market for money.

In a completely specified macroeconomic model, the money and bond markets are only two of the markets in the economy. The other markets are those for commodities and labor. An interdependent structure of such a model implies that the interest rate is jointly determined with the other endogenous variables – including output and price level – of the model. For such a model, Walras's law implies that one of the markets can be omitted from explicit analysis. The loanable funds theory would omit the market for money while the liquidity preference theory would omit the market for bonds from explicit analysis, though both these choices would yield, *ceteris paribus*, the same equilibrium values of all the endogenous variables, including the interest rate. Hence, it does not matter for general equilibrium analysis which of these two theories is adopted in a given macroeconomic model.

Which theory is adopted does matter in a dynamic context. For analyzing the dynamic movements in the interest rate, our preference has been for the theory based on excess demand in the bond market.

However, empirical studies find support for the elements of excess demands for both money and bonds in explaining movement in the interest rate. An essential requirement for such empirical determination of the interest rate is the inclusion of the Fisher equation. While the empirical studies reported in this chapter used a distributed lag model for expectations and found long lags, newer studies tend to use rational expectations. No matter which procedure for modeling expectations is used, most studies report that increases in the money supply decrease the nominal interest rate and that money is not neutral in the short run as far as the nominal and real interest rates are concerned.

Summary of critical conclusions

- ❖ The traditional classical loanable funds theory stated that the real interest rate is determined by full-employment saving and investment in the economy.
- ❖ Keynes argued that there is no direct nexus between saving and investment. His liquidity preference theory asserted that the interest rate is determined by the demand and supply of money.
- ❖ The modern classical theory adapts the traditional classical loanable funds theory to the statement that the real interest rate is determined by the demand and supply ofbonds.
- ❖ The modern classical approach implies that the long-run general equilibrium real rate of interest is invariant to anticipated changes in the money supply and the rate of inflation. Therefore, for the real rate of interest, there is neutrality of anticipated changes in money and inflation.
- ❖ The new Keynesian approach implies that monetary policy can change the real interest rate.
- ❖ Walras's law implies that it is immaterial in general equilibrium whether the money or the bond market is taken to be the proximate determinant of the rate of interest; the rate of interest will be identical.
- ❖ However, dynamic analysis shows that it does matter for the magnitude and in some cases also for the sign of the change in the interest rate whether this change is made a function of the excess demand for money or for bonds.
- ❖ In the modern financially developed economy, the more appropriate *proximate* determinant of changes in the interest rate is the excess demand for bonds. Changes in the excess demand for money cause changes indirectly in the interest rate by first changing the excess demand for bonds.

Review and discussion questions

1. Explain how the monetary and real factors enter into the determination of the interest rates in the short run and in the long run.
2. Compare and contrast the liquidity preference and loanable funds theories of the rate of interest. Discuss their implications for monetary policies intended to maintain full employment.
3. Keynes asserted that there is no such thing as a non-monetary theory of the rate of interest and that the rate of interest is uniquely determined by the demand and supply of money. Explain Keynes's reasons for this view. Compare this view with those of the traditional classical and modern classical schools.

4. Discuss the adjustment process likely to follow a change in (a) the money supply through open market operations, (b) a cut in the central bank's discount rate, leading to eventual changes in the interest rates in the economy.
5. Can the central bank change the interest rate in the economy through changes in its discount/bank rate? Present the analysis and theory relevant to your answer for the economy you live in.
6. Monetary theory implicitly assumes that the interest rates and the money stock are uniquely linked so that changes in one have a corresponding counterpart in the other, so that it does not matter which one the central bank chooses to change. What assumptions are needed for this assertion? Are they realistic enough for policy purposes?
7. This chapter has made the rather unusual assertion that in real-world economies there is no explicit market for money. Instead, the money market is a reflection of the other markets. Do you agree or disagree? Give reasons for your answer. What does your answer imply for the theory relevant to the determination of the interest rate if there is
   (a) general equilibrium in all markets, (b) disequilibrium in the economy?
8. The buffer stock analysis of the demand for money in Chapter 6 asserted that money acts as a buffer during periods in which economic agents need to adjust their stocks of other goods (commodities, bond and labor) to their optimal levels, but that such adjustments are more costly in the short term than those in money balances. What does this imply for the determination of the interest rate if there exists general equilibrium in all markets? What does it imply for the dynamic determination of the interest rate while there are buffer stock holdings of money following a shock that changes the desired demands for other goods?
9. Is there some relationship between the assertions (a) on buffer stock money holdings and (b) that in the real-world economies there is no explicit market in the economy for money but that it is a reflection of the other markets? Discuss.
10. Dynamic adjustments occur in disequilibrium but Chapter 18 raised doubts about the applicability of Walras's law if there was disequilibrium in the commodities and labor markets. In this context, should the dynamic analysis of interest rates be conducted with notional or effective excess demand functions? Discuss, keeping in mind that the objective is to explain the dynamic, disequilibrium determination of the rate of interest.
11. "The assumption of modern classical economics that there exists continuous labor market clearance at full employment means that we can confine the analysis of the real rate of interest to states of general equilibrium and ignore its properties for the disequilibrium states. Therefore, it does not matter whether the loanable funds theory or the liquidity preference theory was used: both imply the same rate of interest by virtue of Walras's law." Discuss the various aspects of this assertion.
12. Do the existence and operations of financial intermediaries have any implications for the rate of interest? If so, are these adequately reflected in the short-run macroeconomic models, and in what ways?
13. "The real rate of interest is a real variable. Under rational expectations, it is invariant to systematic changes in the money supply or the price level. However, unanticipated changes in these nominal variables can change the real rate of interest." Discuss this statement in the context of the neoclassical model and specify

the implied Lucas-type equation for the determination of the real rate of interest. What are its implications for the pursuit of monetary policy?

14. Is there a “natural” rate of interest? What does it mean and what determines it? Is there a curve such as the Phillips curve for the real rate of interest? Discuss.

15. Why does the real interest rate fluctuate over the business cycle? Can monetary factors change it? Discuss.
16. Are the loanable funds and liquidity preference theories of the rate of interest consistent with (i) interest rate targeting, (ii) the Taylor rule? If not, how can they be made consistent?

References

Crowder, W.J., and Hoffman, D.L. “The long-run relationship between nominal interest rates and inflation: the Fisher equation revisited.” *Journal of Money, Credit and Banking*, 28, 1996, pp. 102–18.

Echols, M.E., and Elliot, J.W. “Rational expectations in a disequilibrium model of the term structure.” *American Economic Review*, 66, 1976, pp. 28–44.

Feldstein, M., and Eckstein, O. “The fundamental determinants of the interest rate.” *Review of Economics and Statistics*, 52, 1970, pp. 363–75.

Hume, D. *Of Interest*. 1752. Reprinted in *The Philosophical Works of David Hume.* 4 vols. Boston: Little, Brown and Co., 1854. [Also available at: cepa.newschool.edu/het/profiles/hume.htm].

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Mundell, R.A. “Inflation and real interest.” *Journal of Political Economy*, 71, 1963, pp. 280–3.

Sargent, T.J. “Commodity price expectations and the interest rate.” *Quarterly Journal of Economics*, 83, 1969, pp. 127–40.

Tobin, J. “Money and economic growth.” *Econometrica*, 33, 1965, pp. 671–84.

20 The structure of interest rates

This chapter extends the determination of the single macroeconomic rate of interest to the multitude of interest rates in the economy.

Two of the major reasons for the variations among interest rates are the differences in the term to maturity and the differences in risk. To explain the former, it is important that the riskiness of bonds be held constant across assets of different maturities. This is made possible by confining the comparison to government bonds of different maturities and studying their yield curve. The main theory for explaining the term structure of interest rates is the expectations hypothesis.

Key concepts introduced in this chapter

- ♦ Yield curve
- ♦ Short rate of interest
- ♦ Long rate of interest
- ♦ Expectations hypothesis of interest rates
- ♦ Liquidity premium
- ♦ Segmented market hypothesis
- ♦ Preferred habitat hypothesis
- ♦ Random walk hypothesis

The short-run macroeconomic models of Chapters 13 to 15 have a single (bond) rate of interest, as analyzed in those chapters. However, there is more than one bond interest rate and more than one type of bond in the economy. By definition, the economist's concept of the rate of interest (or yield) on any given asset is the rate of return, including expected capital gains and losses, on that asset over a given period of time. Therefore, there is a rate of interest for each distinct type of asset in the economy. An example of this is provided by Chapter 16, which has two interest rates, one on bonds and the other on credit.

Assets differ in various aspects or characteristics. Some of the more significant differences consist in their marketability, their risk and their term to maturity. The rates of return on assets are likely to differ, depending upon their characteristics. The macroeconomic mode of focusing on only one rate of interest is quite acceptable if all interest rates are related to each other in fixed proportions or fixed differences. Empirically, they do have a high positive correlation. The relationship between prices and rates of return on assets of differing maturities is brought out by the theories on the term structure of interest rates. These theories and the empirical work based on them are the focus of this chapter.

Section 20.1 defines the spot, forward and long rates of interest. Section 20.2 sets out the theories explaining the term structure of interest rates. Of these, the most significant one for developed financial markets is the expectations hypothesis. Section 20.3 briefly touches on the relationship between asset prices and yields, and on tests based on the term structure of asset prices. Sections 20.4 and 20.5 report on some of the empirical work on the term structure of interest rates. Section 20.6 presents the random walk hypothesis which is related to the expectations hypothesis and uses the rational expectations hypothesis for the formation of expectations. Section 20.7 uses the term structure to derive estimates of the expected rate of inflation.

The basic model of the relationship between the prices and yields of assets with different risks is the capital asset pricing model proposed by Sharpe (1964). This analysis is based on the expected utility hypothesis developed in Chapter 5. However, such analysis is usually not included in textbooks on monetary economics and, for reasons of brevity, we have chosen not to include it in this book.

Note that the theories on the risk structure and the term structure of interest rates only explain the interest rate differentials due to differences in risk or the term to maturity, and do not explain the basic interest rate in the economy, which was the subject of Chapter 19.

Notation

Unfortunately, the notation in this chapter has to be quite cumbersome, so that some explanation on its general pattern would be useful. To start, first note that all interest rates in this chapter are nominal. The *short* (nominal, one-period) *interest rates* are designated by r.[1] The current period is designated as t. Suppose that a contract is entered into in period $t \quad j$ for a one-period loan for period $t \quad i$ at an interest rate r, $j \quad i$. This will be written as+ ${}_{t\ j}r_{t\ \ i}$, where the left subscript indic+ates the period in which t≤he contract is made and the +right subscript indicates the period for which the loan is made. If future interest rates are the expected ones, we would write the corresponding rate as ${}_{t+j}r^{e}{}_{t+i}$ and its rational expectation as $E_{t+j}\ {}_{t+j}r_{t+i}$, where the expectation E_{t+j} is based on information available in period $t\ j$+

The *long rates* are designated by the capital symbol R. The contract for these is always assumed to be entered into in the current period t. ${}_{t}\ R_{t+i}$ will designate the long rate on a contract for a loan of i periods. Since this interest rate is known in the current period t, it is an actual rather than an expected rate.

Some of the concepts of the rate of interest

The short-term markets for bonds have spot, forward and long rates of interest. The meanings of these terms are as follows.

[1] Note that the lower-case symbol *r* designates a nominal short rate. In earlier chapters, the nominal rates were designated by the capital symbol *R*. However, in this chapter, *R* will be reserved for the nominal long interest rate.

The (current) spot rate of interest

The (current) spot rate of interest $_t r_t$, or written simply as r_t, is the annualized rate of return on a loan for the current period t, with the loan being made at the beginning of period t.

The future spot rate of interest

The future spot rate of interest is the return on a one-period loan in a future period $(t + i)$, $i > 0$, with the loan made at the beginning of that period. It will be designated $_{t+i}r_{t+i}$ or r_{t+i}, so that the left-hand subscript will be implicit. Since r_{t+i} is a future spot rate, its expected value will be designated $r_{t+i}{}^{e}$. Its rational expectation in period t will, then, be written as $E_t r_{t+i}$, or as $E_{t\ t+i}r_{t+i}$.

The future short rate of interest

The future short rate of interest is the return on a one-period loan in a future period $(t + i)$, $i > 0$, with the contract for the loan entered into at the beginning of period $t + j$, $j \leq i$, which could be the current period. It will be designated $_{t+j}r_{t+i}$.

The forward short rate of interest

The forward short rate of interest $_t r^{f}{}_{t\ i}$ is the annualized rate of interest on a one-period loan for the $(t\ \ i)$th period only, with the contract for the loan being made in the current period t. Note that the superscript f has been inserted to stand for "forward." The forward rate differs from the future short rate $_{t+i}r_{t+i}$ (or r_{t+i}), where the one-period loan for the period $(t + i)$ is contracted at the beginning of period $t + i$. In incomplete financial markets, $_t\, r^{f}{}_{t+i}$ may not exist but $_{t+i}r_{t+i}$ would do so as long as there are spot markets. However, $_t\, r^{f}{}_{t+i}$, if it exists, will be known in the current period t, whereas $_{t+i}r_{t\ \ i}$ is not likely to be known in t, though expectations on its value can be formed in t.

The long rate of interest

The long rate of interest $_t\, R_{t\ i}$, $i\ 0, 1,\ldots, n$, is the rate of return per period on a loan for $(i\ 1)$ periods, the loan being made in period t, with repayment of the principal and accumulated interest after $(i\ 1)$ periods.

The current spot rate of interest $_t r_t$ and the one-period long rate of interest $_t R_t$ are identical. For simplicity of notation, $_t\, r_{t+i}$ will sometimes be written as r_i and $_t\, R_{t+i}$ will be written as R_i, with the subscript t being implicit or with the current period being treated as 0.

Term structure of interest rates

Yield curve

The variation in yields on assets of different maturities (redemption dates) is known as the term structure of interest rates, with the assets being assumed to be identical in all respects except for their maturity. This requirement is generally fulfilled only by the bonds issued by the government, so that the yields on government bonds are examined to show the variation in yield with increasing maturity. This variation is shown graphically by plotting the nominal

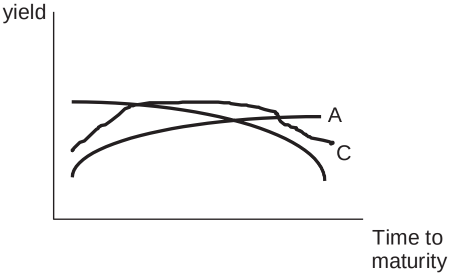


Figure 20.1

yield r on government bonds on the vertical axis and the time up to maturity on the horizontal axis, as in Figure 20.1. The curve thus plotted is known as the *yield curve*.

The yield curve normally slopes upward from left to right, with the yield rising with term to maturity, as shown by the curve A in Figure 20.1. It can, however, possess any shape. In times of monetary stringency, short-term interest rates can rise and move above the long-term rates, as shown by curve B. This can also happen when inflation is rampant in the economy but is expected to be a short-term problem so that the inflationary premium in nominal yields is greater for the shorter term than for the longer term bonds. In some cases, the curve may have a hump, as shown by curve C. In this case, some intermediate securities have the highest yield, usually because of the expectation that the highest rates of inflation will occur in the intermediate periods.

The two main determinants of the shape of the yield curve in practice are the time structure of the expected inflation rates and the current stage of the business cycle. On the former, as explained in several earlier chapters, Fisher's relationship between the nominal yields and the expected inflation rate is:

$$(1 + r_t) = (1 + r^r_t) + (1 + \pi^e_t)$$

where r is now the nominal short yield, r^r is the real short yield and π^e is the expected inflation rate. The higher the expected rate of inflation, the more will the time structure of expected inflation determine the shape of the yield curve.

The yield curve changes its shape over the business cycle. Long-term yields are usually higher than short-term yields mainly because long-term debt is less liquid and is subject to greater price uncertainty than short-term debt. However, the short-term yields are more volatile, rising faster and extending further than long yields during business expansions and falling more rapidly during recessions. Large swings in short-term rates, and to a lesser extent in intermediate rates, together with relatively narrow movements in long-term rates, cause a change in the shape of the yield curve over the course of a business cycle.

A sharp increase in short-term rates frequently occurs near the peak of a business expansion because of a combination of factors, most often including a strong demand for short-term credit, restrictive effects of monetary policies on the supply of credit, and changing investor expectations. Depending upon the intensity of these forces, the yield curve will be relatively flat, have a slight downward slope, or show a steep negative slope. As short rates fall absolutely and relative to long yields during the ensuing economic slowdown, the yield curve tends to regain its positive slope, acquiring its steepest slope near the cyclical trough. As the economy recovers and economic activity picks up, short rates again rise faster than long yields, and the yield curve tends to acquire a more moderate slope. Since the yield curve plots the nominal rather than the real rate of interest, and the nominal rate includes the expected rate of inflation, the dominant element of the shape and shifts in the yield curve is often the term structure of the expected rate of inflation.

There are basically three main theories on the term structure of interest rates. These are:

1 The expectations hypothesis, first formulated by Irving Fisher. This theory is the relevant one for financially developed markets, and is supported by most empirical studies.
2 The segmented markets theory, with Culbertson as its major proponent.
3 The preferred habitat hypothesis.

Expectations hypothesis

Irving Fisher in *The Theory of Interest* (1930, pp. 399–451) considered the rate of return or yields on securities that differ only in terms of their maturity. His approach assumes that:

(i) All borrowers and lenders have perfect foresight and know future interest rates and asset prices with certainty, so that there is no risk. An alternative assumption to this is that, while there is uncertainty of yields, the borrowers and lenders are risk neutral and form rational expectations about the future short rates.
(ii) There are no transactions costs in switching from money into securities and vice versa.
(iii) The financial markets are *efficient*.

A market is said to be *efficient* if it clears (i.e. demand equals supply) instantly and prices reflect all available information. In such a market, any opportunities for superior profits are instantly eliminated. By comparison, a *perfect market* assumes perfect competition among traders *and* efficient markets. Fisher's assumptions specify an efficient market, which need not have perfect competition, so that it need not be a perfect one.

Investors are assumed to maximize their expected utility, subject to the relevant constraints. However, under assumption (i), this is synonymous with the maximization of the expected return to the portfolio. Under assumptions (i) and (ii), a lender wishing to make a loan for *n* periods will be indifferent between an *n*-period loan or a succession of *n* one-period loans only if the overall return were the same in both cases. Under assumption (iii), with all investors acting on this basis, the market yields will be such as to ensure this indifference.

Expectations hypothesis, complete markets and forward rates

Assume that the financial sector has complete markets, so that there exist markets for long loans of all possible maturities, as well as for spot and *forward* one-period loans. With the current period as *t*, the yield (per period) on an $(i + 1)$-*period* loan was designated as ${}_tR_{t+i}$, while that on a *one-period* loan for the $(i + 1)$th period was ${}_tr^f$, $i = 0, 1, \ldots, n$, where $n +$ 1 is the longest maturity available in the market. Hence, r $_{t+i}$

$_{t\ t}$ is the (spot) yield on a loan for the first period; ${}_tr^f{}_{t+1}$ is the forward yield on a loan for the second period; and so on. An $(i + 1)$-*period* loan of \$1 will pay the lender $(l + R_{t+i})^{i+1}$ at the end of the $(i + 1)$th period.

The series of $(i+1)$ loans starting with a principal of $1 for one-period at a time will pay him $[(1+r_t)(1+r^f_{t+1})\ldots(1+r^f_{t+i})]$ at the end of the $(i+1)$th period. Under the above three assumptions, the lender will be indifferent between the two types of loans if the total amount repaid to him after $n+1$ periods is identical. With all investors exhibiting this behavior, efficient markets under certainty ensure that:

$$(1+{}_tR_{t+i})^{i+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2})\ \ldots\ (1+{}_tr^{\mathrm{f}}{}_{t+i}) \tag{1}$$

This formula will hold for every i, $i = 0,\ldots, n$, where $n+1$ is the longest maturity in the market, so that:

$$(1+{}_tR_t) = (1+{}_tr_t)$$

$$(1+{}_tR_{t+1})^2 = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})$$

$$(1+{}_tR_{t+2})^3 = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2})$$

....

$$(1+{}_tR_{t+n})^{n+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2})\ \ldots\ (1+{}_tr^{\mathrm{f}}{}_{t+n}) \tag{2}$$

Under our assumption of complete markets, the forward rates are known, rather than merely expected, in period t. However, even well developed financial markets do not have forward markets for all future periods, so that (2) cannot be applied for all maturities.

Expectations hypothesis and expected future spot rates

Since there would always be spot markets over time, designate the spot rate expected in period t for the period $t+i$ as ${}_tr^{\mathrm{e}}{}_{t+i}$, where the subscript t on the left side in the presence of the superscript e indicates that the expectations are formed in period t for the spot rate for period $t+i$.[2] The investor would then have a choice of investing long for $t+i$ periods, with a known long rate ${}_tR_{t+i}$, and investing over time in a sequence of spot markets at the spot rates in those markets. In practice, since these future spot rates can differ from the actual ones, there is a risk in following the latter strategy. The investor will be indifferent between the two strategies if he is *risk indifferent* and if their expected return is identical. Hence, in terms of the expected future rates, the expectations hypothesis becomes:

$$(1+{}_tR_{t+i})^{i+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{e}}{}_{t+1})(1+{}_tr^{\mathrm{e}}{}_{t+2})\ldots\ (1+{}_tr^{\mathrm{e}}{}_{t+i}) \tag{3}$$

Note that (3) differs from (1) since (3) involves expected future spot rates while (1) involves the corresponding forward rates, which are known in period t. For many investors, though ones with relatively small portfolios, the assumptions of the expectations hypothesis can be somewhat unrealistic. There is often both a transfer cost in and out of securities and a lack of perfect foresight (or risk indifference) about the future. The former implies that n one-period loans will involve much greater expense and inconvenience than a single n-period loan. The latter implies that loans of different maturities involve different risks and, for risk averters, a higher risk has to be compensated for by a higher yield. For very many large transactors, usually financial institutions, the transactions costs tend to be negligible, so that (3) should hold approximately, if not accurately.

[2] Note that ${}_t r^{f}_{t+i}$ is a forward rate contracted in t for the one-period loan in $t + i$, whereas ${}_t r^{e}_{t+i}$ is the spot rate for

$t + i$ with expectations formed in t.

Under the rational expectations hypothesis, r^e is replaced by $E_t r$, so that (3) becomes:

$$(1 +_t R_{t+i})^{i+1} = (1 +_t r_t)(1 + E_{t\,t} r_{t+1})(1 + E_{t\,t} r_{t+2}) \ldots (1 + E_{t\,t} r_{t+i}) \qquad (3^J)$$

If a difference emerges in the markets between the left and the right sides of (1) and (3), profits can be made through arbitrage, which would take place to establish their equality. The rest of this chapter proceeds in terms of (3) or (3^J) rather than (1). While financial markets, even in developed economies, rarely have a large number of forward markets, they usually do have markets for government securities of many different maturities. The long rates of interest are quoted on these securities, so that their values are known each period. These values can be used to calculate the expected short rates of interest by using the following iterative reformulation of (3):

$$E_{t\,t} r_{t+1} = (1 +_t R_{t+1})^2/(1 + {}_t r_t) - 1$$

$$E_{t\,t} r_{t+2} = (1 +_t R_{t+2})^3/[(1 + {}_t r_t)(1 + E_{t\,t} r_{t+2})] - 1 \qquad (4)$$

and so on.

If the market forms its expectations in terms of the expected future short rates, the long rates will be determined from these short rates by the preceding equations. Some economists assume that the investors' expectations are formed in terms of a series of expected short rates for the future periods, while others assume that investors are concerned with the prices of the assets currently in the market and that these prices can be used to calculate the long rates. Therefore, equation (3) can be used from right to left or from left to right.

Long rates as geometric averages of short rates

According to (3), the long rates are geometric averages of the short rates of interest. This implies that:

1 If the short interest rates are expected to be identical, the long rate will equal the short rate.
2 If the short interest rates are expected to rise, the long rates will lie above the current short rates.
3 If the short interest rates are expected to decline, the long rates will be less than the current short rate.
4 The long rate, being an average of the short rates, will fluctuate less than the short rate.

In principle, any pattern of expected future short rates is possible, with the result that some long rates may be less and some greater than the current spot rate, so that the yield curve may have any shape whatever.

The assumptions of the expectations hypothesis may not always hold for all agents in the market, which encompasses both households and firms. However, developed financial economies tend to be competitive and efficient. Therefore, the expectations hypothesis will hold if the credit markets have sufficient numbers of participants who behave according to the assumptions of perfect foresight (or of rational expectations and risk indifference) and zero variable transfer costs between securities and money. These assumptions tend to be valid at least for large financial institutions operating in the developed economies. Hence, the expectations hypothesis should be more or less valid for developed financial markets.

Once the market has established a structure of short and long rates according to (1) or (3), the demand and supply functions for long and short bonds on an individual basis will become *indeterminate*: an investor would be indifferent between a long bond maturing at the end of the *i*th period and various sequences of short and long bonds with a corresponding combined maturity.

Liquidity preference version of the expectations hypothesis

Both the *n*-period loan and a series of *n* one-period loans involve risks, though of different kinds. The *n* one-period loans involve the possibility that the future spot rates will turn out to be lower than the expected forward rates or the *n*-period long rate. This is an *income* loss. But the *n*-period loan – that is, purchase of a bond maturing after n periods – involves the possibility that the lender may need his funds somewhat sooner and have to sell the bond before it matures. Such a sale may involve a *capital* loss, especially in the absence of a secondary market for loans. There is also the possibility that more profitable opportunities may turn up and have to be foregone if the funds are already loaned up for a long period.

It is likely that the possibility of a capital loss influences lenders' decisions more than that of the interest loss since the capital loss can usually take on much greater magnitude than the interest loss. Further, if the funds represent precautionary saving, the individual would prefer a more liquid (shorter maturity) to a less liquid (longer maturity) asset. Hicks (1946, pp. 151–82) suggested that lenders wish to avoid the risk of a capital loss by investing for shorter rather than longer periods. Therefore, under uncertainty of future yields, they have to be compensated by a higher yield on longer term loans. Conversely, borrowers – generally firms borrowing for long-term investments – prefer borrowing for a longer term than for a shorter term, which makes them willing to pay a premium on longer term loans. Such risk-avoidance behavior on the part of both lenders and borrowers implies that the longer term loans will carry a premium over shorter term loans. Hence, the yield on bonds will increase with the term to maturity, so that equation (3) will be modified to:

$$(1 + {}_tR_{t+n})^{n+1} > (1 + {}_tr_t)(1 + {}_tr^{\mathrm{e}}{}_{t+1})\ldots(1 + {}_tr^{\mathrm{e}}{}_{t+n}) \qquad n \geq 1 \qquad (5)$$

Equation (5) is known as the *liquidity preference hypothesis of the yield curve*. For a more specific hypothesis on liquidity preference, designating the *liquidity premium* as ${}_t\gamma_{\underset{+}{t}\ n}$, we have:

$$(1 + {}_tR_{t+n})^{n+1} = (1 + {}_tr_t)(1 + {}_tr^{\mathrm{e}}{}_{t+1})\ldots(1 + {}_tr^{\mathrm{e}}{}_{t+n})\,{}_t\gamma_{t+n}(n;\rho) \qquad n \geq 1 \qquad (6)$$

where $\partial\gamma_n/\partial n \geq 0$ by virtue of the liquidity premium, and:

$\gamma =$ liquidity premium
$\rho =$ degree of risk aversion
$n =$ periods to maturity.

We can distinguish between two versions of (6) on the basis of two alternative assumptions on the liquidity premium. These are that:

(i) The liquidity premium is constant at γ per period, so that ${}_{t+}\gamma_{t=}i\gamma$. While there is no particular intuitive justification for making this assumption, it is analytically convenient and, as seen later in this chapter, is made in many empirical studies. It reduces (6) to:

$$(1 + {}_tR_{t+n})^{n+1} = (1 + {}_tr_t)(1 + {}_tr^{\mathrm{e}}{}_{t+1}) \ldots (1 + {}_tr^{\mathrm{e}}{}_{t+n})n\gamma \qquad n \geq 1 \qquad (7)$$

Equation (7) with a constant per period risk premium is sometimes called the *strong form* of the expectations hypothesis with a liquidity premium.

(ii) The per period liquidity premium varies with the term to maturity and, moreover, may not be constant over time, e.g. over the business cycle, so that (6) does not simplify to (7). This is sometimes called the *weak* form of the expectations hypothesis with a liquidity premium. Estimation of this form requires specification of the determinants of the liquidity premium.

Compared with these *weak* and *strong* forms of the expectations hypotheses, the original form (3) of this hypothesis without a liquidity premium is known as the *pure* form of the expectations hypothesis.

Segmented markets hypothesis

If the uncertainty in the loan market is extremely severe or if lenders and borrowers have extremely high risk aversion, each lender will attempt to lend for the exact period for which he has spare funds and each borrower will borrow for the exact period for which he needs funds. In this extreme case, the overall credit market will be split into a series of segments or separate markets based on the maturity of loans, without any substitution by either borrowers or lenders among the different markets. Therefore, the yields in any one market for a given maturity cannot influence the yields in another market for another maturity. Hence, there would not be any particular relationship such as (3) or (6) between the long and the short rates, and the yield curve could have any shape whatever. This is the basic element of the segmented markets theory: the market is segmented into a set of independent markets. Culbertson (1957) stressed this possibility as a major, though not the only, determinant of the term structure of interest rates.

Culbertson also argued that the lender rarely knows in advance exactly when he will need his funds again and will prefer to make loans for shorter terms rather than longer ones, the former being the more liquid of the two. If the supply of short-term debt instruments is not sufficient to meet this demand for liquidity at a rate of interest equal to the long-term rate, the short-term rate will be less than the long-term rate. Further, the supply of short-term instruments is generally limited since lenders will not finance long-term investment with short-term borrowing. Therefore, the short-term yield will be less than the long-term yield, *ceteris paribus*.

The segmented markets hypothesis is more likely to be applicable in the absence of developed financial markets, including secondary markets for securities, and sophisticated investors. It may therefore be somewhat more valid for financially underdeveloped markets than for developed ones.

Preferred habitat hypothesis

The preferred habitat hypothesis was proposed by Modigliani and Sutch (1966, 1967), and represents a compromise between the expectations hypothesis of perfect substitutability and the segmented markets hypothesis of zero substitution between loans of different maturities. Modigliani and Sutch argued that lenders would prefer to lend for periods for which they can spare the funds and borrowers would prefer to borrow the funds for periods for which they need the funds. However, each would be willing to substitute other maturities, depending upon their willingness to take risks and the opportunities provided by the market to transfer easily among different maturities. Bonds maturing close together would usually be fairly good substitutes and have similar risk premiums. This would be especially so for bonds at the longer end of the maturity spectrum. Therefore, in well developed financial markets, a high degree of substitutability would exist among different maturities, but without these necessarily becoming perfect substitutes. Hence, while the yields on different maturities would be interrelated to a considerable extent, there would also continue to exist some variation in yields among the different maturities.

Implications of the term structure hypotheses for monetary policy

The expectations theory and the segmented markets theory have significantly different implications for the management of the public debt and for the operation of monetary policy. The expectations theory implies that the market substitution between bonds of different maturities is so great that a shift from short-term to long-term borrowing by the government will not affect the shape of the yield curve. The segmented markets theory implies that a substantial purchase (sale) of short-term bonds will lower (raise) the short-term interest rates while a sale (purchase) of long-term bonds will raise (lower) the long-term rates, so that such policies can alter the yield curve. The implications of the preferred habitat hypothesis lie between those of the expectations hypothesis and the segmented markets hypothesis, and are closer to one or the other depending upon the stage of development of the financial markets and the characteristics of the economic agents operating in them.

The empirical evidence for economies with well-developed financial markets has so far generally favored the expectations theory or a version of the preferred habitat hypothesis close to the expectations hypothesis over the segmented markets hypothesis. Intuitively, the credit markets for such economies are not seriously segmented since borrowers and lenders do generally substitute extensively between assets of different maturities.[3] A number of studies for the USA and Canada have substantiated the expectations theory at the general level, though there also exist many empirical studies that reject its more specific formulations. We discuss some of these later in this chapter.

Financial asset prices

Financial assets are not generally held for their direct contribution to the individual's consumption. They are held for their yield, which is often uncertain, and the individual balances the expected yield against the risks involved. This is the basic approach of the theories of portfolio selection. These theories focus on the yields on assets rather than on the prices of assets.

The price of any asset is uniquely related to its yield and can be calculated from the following relationship. In any period *t*, for an asset *j*,

$$r_{jt} = ({}_tp^{\mathrm{e}}{}_{jt+1} - p_{jt}) + x_{jt} \qquad (8)$$

[3] An early study by Meiselman in 1962 supported the expectations theory. He also found that there was not sufficient justification for the assumption of a liquidity premium.

where:

r_{jt} = expected yield on the *j*th asset during period *t*
p_{jt} = *j*th asset's price in period *t*
${}_tp^e{}_{j\,t+1}$ = *j*th asset's (expected) price in period $t + 1$, with expectations held in *t*
x_{jt} = *j*th asset's coupon rate in period *t*.

That is,

$$ {}_tp^e{}_{jt+1} = p_{jt} + r_{jt} - x_{jt} \tag{9}$$

Hence, a theory of the rate of interest is also a theory of the prices of financial assets. Alternatively, the yields on assets may be explained by a theory of asset prices. Such a theory at a microeconomic level would consider the market for each asset, and use the demand and supply functions for each asset to find the equilibrium price of the asset. At the macroeconomic level, the theory could focus on the average price of financial assets, with macroeconomic demand and supply functions. These demand and supply functions would have the prices of the assets as the relevant variables. This suggests two structural estimation procedures. One of these would specify the demand and supply functions for financial assets in terms of their prices, while the second one would do so in terms of interest rates. The former procedure would derive the equilibrium prices of assets with different maturities, which can then be used to calculate the short and long interest rates. The latter procedure would derive the equilibrium short and long interest rates, which can be used to calculate the prices of assets of different maturities. These procedures are not explicitly specified in this book but can be found in its first edition (2000). Empirical studies based on these structural approaches include, among others, Benjamin Friedman (1977), Feldstein and Eckstein (1970), Sargent (1969) and Echols and Elliot (1976).

An illustration of the arguments and findings of such studies is provided by Echols and Elliot (1976). These authors extended Sargent's (1969) analysis and tested for the determinants of forward rates using real GNP, government deficit, net export balance, real money supply, stock of outstanding government bonds, bank funds and insurance company funds invested in government bonds – as well as inflationary expectations – among their explanatory variables. Their estimations of forward rates for US data found the coefficients of the explanatory variables to be significant, with signs consistent with the loanable funds approach. Among their results was the significance of the liquidity premium, as well as that of the institutional (bank and insurance company) demands for bonds of different maturities and the supply of short versus long maturity supplies of government bonds, thereby supporting the preferred habitat hypothesis.[4] For example, an increase in the proportion of investment funds held by banks relative to insurance companies lowered forward rates. However, these institutional holdings and supply factors did not prove to be significant in explaining the yield spread between twelve-year government bonds and Treasury bills, so that it is not clear that the government could shift the yield curve by debt management policy – e.g. by increasing the issue of short-term government bonds relative to long-term bonds.

Compared with these structural approaches, the ones usually employed to test the theories of the term structure of interest rates are reduced-form approaches. These are based on the expectations hypothesis and are considered in detail below.

[4] These results of Echols and Elliot (1976) should be compared with those in Pesando (1978), discussed later in this chapter.

Empirical estimation and tests

Reduced-form approaches to the estimation of the term structure of yields

As shown earlier, the expectations hypothesis modified with the addition of a liquidity preference term implied (6), which was that:

$$(1 + {}_tR_{t+n})^{n+1} = [(1 + {}_tr_t)(1 + {}_tr^{e}_{t+1}) \ldots (1 + {}_tr^{e}_{t+n})]\,{}_t\gamma_{t+n}(n;\rho) \qquad n \geq 1 \qquad (10)$$

where the current period is *t*, the expectations are those held in period *t* and γ represents the liquidity premium, which depends on the term to maturity *n* and the risk premium ρ. Now, using the symbols *R* and *r* for the *logarithmic* values of the *gross* (rather than the net) returns and adding a random error η_t, the estimation form of (10) becomes:

$${}_tR_{t+n} = \{1/(n+1)\}[{}_tr_t + {}_tr^{e}_{t+} + \ldots + {}_tr^{e}_{t+} +_t \gamma_{t+n}(n;\rho)] + \eta_t \qquad n \geq 1 \qquad (11)^5$$

where *all the return variables are gross rates of return and all symbols now indicate log values*. We will follow this convention in the rest of this chapter.

In order to test (11), we need a hypothesis for generating the expected values of the forward short rates. Given the efficient markets assumption, the natural complement of the expectations hypothesis is the rational expectations hypothesis (REH) presented in Chapters 8 and 14. This hypothesis specifies that:

$${}_tr^{e}_{t+i} = E_t r_{t+i} \qquad (12)$$

where $E_t\, r_{t\ i}$ is the rational expectations value of $r_{t\ i}$ based on all the information available in *t* about period *t i*. Among the information needed is that of the "relevant theory" that actually determines $+r_{t\ i}$ and information on the values of its determinants. This throws us back to the demand and supply functions for assets for the (*t i*)th period. Alternatively, if we can assume that this theory and the values of the explanatory variables are all known, we can estimate the stochastic equation (11).

As an illustration of this point, assume that the *relevant theory* is the simple autoregressive relationship with a stochastic term:

$$r_{t+i+1} = a_1 r_{t+i} + a_2 r_{t+i-1} + \mu_{t+i+1} \qquad i = 0, 1, 2, \ldots, n \qquad (13)$$

where μ_t is a random error with zero mean and constant variance. Under the REH,

$$E_t r_{t+i+1} = a_1 E_t r_{t+i} + a_2 E_t r_{t+i-1} \qquad (14)$$

where $E_t\, r_{t+i}$ is the rational expectation of the expected spot rate ${}_tr^{e}_{t+i}$, with the expectations formed in *t*. By iteration, $E_t\, r_{t\ i}$ can be expressed as functions of r_t and $r_{t\ 1}$, whose values are already known in *t*. These, along with the expectations hypothesis of the term structure, can be used to generate $E_t\, r_{t+i}$ for all *i*. The following provides an example of these arguments for $i = 2$.

[5] Some researchers have called (11) the “*fundamental equation*” of the term structure and bond pricing.

For our example, (11) fori=2 (that is, three period) becomes:

$$ {}_tR_{t+2} = (1/3)[r_t + {}_tr^e_{t+1} + {}_tr^e_{t+2} + {}_t\gamma_{t+2}] + \eta_t \quad (15)$$

Combining (15) with the REH, we get:

$$E_tR_{t+2} = (1/3)[r_t + E_tr_{t+1} + E_tr_{t+2} + {}_t\gamma_{t+2}] \quad (16)$$

where $E_t R_{t+2}$ is the mathematical expectation in t of the long rate ${}_tR_{t+2}$ from t to $t+2$. From (16) and (14), we have:

$$E_t R_{t+2} = (1/3)[r_t + (a_1r_t + a_2r_{t-1}) + (a_1^2 + a_2)r_t + a_1a_2r_{t-1}] + {}_t\gamma_{t+2}$$

$$= \alpha_1r_t + \alpha_2r_{t-1} + (1/3){}_t\gamma_{t+2} \quad (17)$$

where $\alpha_1 = (1/3)(1 + a_1 + a_2 + a_1^2)$ and $\alpha_2 = (1/3)(a_2 + a_1a_2)$. Since the REH implies that:

$$ {}_tR_{t+2} = E_tR_{t+2} + \eta_t \quad (18)$$

we have from (17) and (18) that:

$$ {}_tR_{t+2} = \alpha_1 r_t + \alpha_2 r_{t-1} + (1/3){}_t\gamma_{t+2} + \eta_t \quad (19)$$

(19) is the estimating equation given the expectations hypothesis, the REH and the assumed specification of the relevant theory as (13). Its general form is:

$$ {}_tR_{t+i} = \alpha'_1 r_t + \alpha'_2 r_{t-1} + (1/(i + 1)){}_t\gamma_{t+i} + \eta_t \quad (20)$$

for appropriate definitions of α'_1 and α'_2 in terms of a_i. Their estimated values would reflect the influence of the three underlying hypotheses. However, note that (20) requires data on the risk/liquidity premium ${}_t\gamma_{t+i}$, which is not observable, so a hypothesis on it will have to be specified before (20) can be estimated. The usual assumptions on ${}_t\gamma_{t+i}$ are considered next.

Two common hypotheses on the risk premium

The simplest possible hypotheses on the risk/liquidity premium ${}_t\gamma_{t+i}$ are:

(i) ${}_t\gamma_{t+i}$ is constant per period such that ${}_t\gamma_{t+i} = i\gamma$, where the liquidity premium for ($i+1$) periods involves this premium for only i periods (after the current one). In this case, this term will become the constant in (20).

(ii) ${}_t\gamma_{t+i}$ is random such that ${}_t\gamma_{t+i} = \xi_{t+i}$. In this case, the liquidity term will become part of the random term in (20).

(i) is the more common assumption in the estimation of (20) and is used in the next section.

Tests of the expectations hypothesis with a constant premium and rational expectations

There is no particular basis for assuming that the liquidity premium is constant. However, making such an assumption facilitates the construction of empirical tests of the expectations hypothesis. The following two tests are based on this assumption. These tests use the implications of the expectations hypothesis with a constant liquidity premium, so that the expected (holding period) yields on the relevant bonds of different maturities will differ only by a constant representing the liquidity premium.

Define the difference between the *actual* long yield and the *average* one specified by the right side of the expectations equation (1) as the excess yield on the long bond. From (11), the actual difference in the yields from holding a long bond as opposed to a sequence of short bonds is related to the liquidity premium and expectational errors. Assuming this premium to be constant and assuming rational expectations, the remaining variations in the excess yields can then only be due solely to random fluctuations. This, in turn, implies that the difference between the excess yield and the premium will be due solely to random errors in expectations and cannot be forecast with any information known at the time the expectations are formed.

Slope sensitivity test

For this test, start with the following non-stochastic form for the *two*-period long rate:

$$2\,{}_tR_{t+1} = {}_tr_t + {}_{t+1}r^{\mathrm{e}}{}_{t+1} + {}_t\gamma_{t+1} \qquad t = 0, 1, \ \ldots \tag{21}$$

where, as a reminder, note that all the variables are in logs and the interest rate variables are gross rates. Assuming the constancy of the liquidity premium per period, and noting that a two-period loan involves a liquidity premium only for the second period, let:

$${}_t\gamma_{t+1} = \gamma \tag{22}$$

(21) can be restated as:

$${}_{t+1}r^{\mathrm{e}}{}_{t+1} - {}_tr_t = 2[{}_tR_{t+1} - {}_tr_t] - \gamma \tag{23}$$

Assuming the rational expectations hypothesis,

$${}_{t+1}r^{\mathrm{e}}{}_{t+1} = E_t\,{}_{t+1}r_{t+1} \tag{24}$$

and

$${}_{t+1}r_{t+1} = E_t\,{}_{t+1}r_{t+1} + \mu_{t+1} \tag{25}$$

where μ_t is a random error with $E_t\,(\mu_{t+1}\,|I_t\,) = 0$ and I_t is the information available in t. Hence, from (23) to (25),

$${}_{t+1}r_{t+1} - {}_tr_t = \alpha + \beta({}_tR_{t+1} - {}_tr_t) - \mu_{t+1} \tag{26}$$

where $\alpha = \ \ \gamma$ and $\beta \ \ \ 2$. Since each of the variables (except for the random term) in (26) is observable, it can be estimated by the appropriate regression technique. This equation specifies that the change over time in the one-period spot rates from the current to the next period will depend upon the difference between the current two-period long rate and the current one-period spot rate, except for a constant term and a random term. New information appearing in the next period will do so randomly. If the regressions of the above equation yield estimates consistent with these restrictions on α, β and μ, the theory will not be rejected by the data.

Equation (26) provides a joint test of three hypotheses: the expectations hypothesis, the REH and a constant liquidity premium per period. Since the test is that the estimated value of β does not significantly differ from 2, it is called the *slope sensitivity test*.

The slope sensitivity test is among the more common ones used for expectations hypotheses. To illustrate this application, this test was used by Mankiw and Miron (1986), among others. They tested equation (26) for the United States using three month and six month data for five intervals during 1890–1979. The null hypothesis (β= 2) was rejected for all except the earliest period prior to the founding of the Federal Reserve System in 1915. That is, the spread between the short and the long rate was a good indicator of the path of interest rates prior to the commencement of the stabilization operations of the Fed, but not in the periods after it, with the spot rate following a random walk after 1915. Therefore, the authors concluded that a central bank policy of interest rate stabilization would make the spot rate follow a random walk and lead to a rejection of the expectations hypothesis. In general, there seems to be more empirical support for (26) when countries do not pursue interest rate stabilization policies.

Efficient and rational information usage test

Another test of the above joint hypothesis is based on the following restatement of (23):

$$_t\varphi^{e}{}_{t+1} \equiv 2\ {}_tR_{t+1} - {}_{t+1}r^{e}{}_{t+1} - {}_tr_t = \gamma \tag{27}$$

where $_t\varphi^{e}{}_{t+1}$ is the *excess yield* over two periods, which, under the constant liquidity premium version of the expectations hypothesis, equals γ . AssumingREH,

$$E_t\,{}_t\varphi_{t+1} = 2\ {}_tR_{t+1} - E_t\,{}_{t+1}r_{t+1} - {}_tr_t = \gamma \tag{28}$$

The stochastic form of this equation is:

$$\begin{aligned} {}_t\varphi_{t+1} &= 2\ {}_tR_{t+1} - {}_{t+1}r_{t+1} - {}_tr_t + \mu_{t+1} \\ &= \gamma + \mu_{t+1} \end{aligned} \tag{29}$$

where $\mu_{t\ 1}$ is again a random term with $E_t\,(\mu_{t\ \ 1}\ I_t\ )=\ 0$, and I_t is the information available in t. Under the joint hypothesis, the excess yield would not be a function of information known in period t. If a regression of the excess yield on information known in t – such as on prices, output, unemployment, and other variables on which information is commonly available in t – yields significant coefficients for such variables, the joint hypothesis will be rejected by the data. From (29), the regression equation can be formulated as:

$$_t\varphi_{t+1} = \alpha + \boldsymbol{bX}_t + \mu_{t+1} \tag{30}$$

where $\alpha\ \gamma$, $\boldsymbol{X}_t$ is a vector of commonly known variables in t and $\boldsymbol{b}$ is the corresponding vector of coefficients. Among the $\boldsymbol{X}$ variables would be included the lagged values of the excess yield itself. The maintained hypothesis would be rejected if any of the estimated coefficients in $\boldsymbol{b}$ were significantly different from zero.

Alternatively, in (29), define:

$${}_t\varphi^{J}_{t+1} = 2\,{}_tR_{t+1} - {}_{t+1}r^{e}_{t+1} \tag{31}$$

and specify the regression equation as:

$${}_t\varphi^{J}_{t+1} = \alpha + \beta\,{}_tr_t + \boldsymbol{bX}_t + \mu_{t+1} \tag{32}$$

where β 1. The joint hypothesis would be rejected if the estimated value of β were significantly different from one and/or if any of the *b* coefficients were significantly different from zero. Jones and Roley (1983) tested (32) for quarterly US Treasury bill data for the period 1970 to 1979. The coefficients of some of the X_t variables were significant, so that the joint hypothesis was rejected.

Many studies using the notion of already available information to test the joint hypothesis for changes in the term structure tend to reject it.[6] A rejection of the joint hypothesis could be due to rejection of the expectations hypothesis, of the assumption of a constant liquidity premium, of the REH, of the proxy used for the expected rate of inflation, or of any combination of them. Therefore, it is not clear whether the expectations theory itself is at fault, since the rejection of the joint hypothesis is sometimes interpreted as a rejection of the assumption of a constant risk premium, sometimes of the REH and sometimes of the proxy used for the expected inflation rate.

The rejection of the assumption of the constancy of the liquidity premium per period implies that this premium can vary over time. This is not implausible for bonds of medium or long maturity, but there is no particular reason to assume that the liquidity premium would vary significantly over periods as short as a week or a few months. Since many of the rejections of the joint hypothesis occur for data using Treasury bill yields only, such rejection may be due to that of the REH or the expectations hypothesis itself. Further, if the liquidity premia are not constant, then the theory needs to specify their determinants, which is difficult to do.

Random walk hypothesis of the long rates of interest

Start with equation (10). With *R* and *r* now redefined as gross rates of interest, (10) becomes:

$$\begin{aligned}{}_tR_{t+n} = {} & (1/(n+1))[r_t +{}_t r^{e}_{t+1} +{}_t r^{e}_{t+2} + \cdots +{}_t r^{e}_{t+n}] \\ & + (1/(n+1))_t\gamma_{t+n}(n;\rho) \qquad n \geq 1\end{aligned} \tag{33}$$

Lagging (33) by one-period:

$$\begin{aligned}{}_{t-1}R_{t+n-1} = {} & (1/(n+1))[r_{t-1} +{}_{t-1}r^{e}_{t} +{}_{t-1}r^{e}_{t+1} + \cdots +{}_{t-1}r^{e}_{t+n-1}] \\ & + (1/(n+1))\,{}_{t-1}\gamma_{t+n-1}(n;\rho)\end{aligned} \tag{34}$$

[6] However, Pesando (1978) reported support for it in Canadian data. This study is discussed in the next section.

Subtract (34) from (33). Applying the REH to the resulting equation gives:

$${}_tR_{t+n} - {}_{t-1}R_{t+n-1} = (1/(n+1))[(r_t - E_{t-1\,t-1}r_t) + (E_{t\,t}r_{t+1} - E_{t-1\,t-1}r_{t+1})] + \cdots$$

$$+(E_{tt}\,r_{t+n-1}) - E_{t-1\ t-1}\,r_{t+n-1})] + (1/(n+1))[(E_{tt}\,r_{t+n} - r_{t-1})]$$

$$+(1/(n+1))[{}_t\gamma_{t+n}(n;\rho) - {}_{t-1}\gamma_{t+n-1}(n,\rho)] \qquad (35)$$

Assuming that no new information becomes available in period t – that is, I_t I_{t-1}, where I_t is the information available in t – we have: =

$$E_{t\,t}r_{t+i} - E_{t-1\,t-1}r_{t+i} = \mu_{t+i} \qquad i = 0, 1, \ldots, n-1 \qquad (36)$$

where $\mu_{t\ i}$ are forecasting random errors with a zero mean and are independently distributed. That is, the revisions to expectations are zero-mean independent random variables.

Further, as $n \to \infty$,

$$(1/(n+1))[(E_t\,r_{t+n} - r_{t-1})] \to 0 \qquad (37)$$

so that, for large n, this term would be about zero for the usually observed and expected range of values of the interest rate. Hence, if the liquidity premium term on the right-hand side was also a random term or if it equaled zero, ${}_tR_{t\ n\ t\ 1}R_{t\ n\ 1}$ would behave randomly. Noting that the last term on the right-hand side of (35) involves the difference between the n-period liquidity premiums in t and t 1, an assumption that the liquidity premium is time invariant –that is, does not change with new information – would make this term equal to zero. Alternatively, it can be assumed that, in the absence of any new information, the last term in (35) will also be randomly distributed.

Hence, under the above collection of assumptions, the right hand side of (35) will be a random variable, so that it can be rewritten as:

$${}_tR_{t+n} = {}_{t-1}R_{t+n-1} + \varepsilon_t \qquad (38)$$

where ε_t is a random error, made up of the relevant set of random errors, with $E(\varepsilon_t\ I_{t\ 1})\quad 0$. (38) states that for large values of n the long rates will follow a random walk. This constitutes the random walk hypothesis (RWH) of the long interest rates. Note that it is expected to hold only for large values of n and if no new information becomes available between t and $t - 1$.

Since systematic monetary policy followed in t will be anticipated in $t - 1$, (38) implies that it cannot affect the change in the long rate between periods $t + n$ and $t + n -$ 1. Only unanticipated monetary policy – that is, policy shocks which change the value of ε – can shift this difference and shift the yield curve. Hence, the RWH of the long interest rates implies that systematic monetary policy cannot shift the yield curve; only unanticipated monetary policy can do so.

However, since we needed (37) to arrive at (38), note that the RWH does not hold for low values of n. To illustrate the failure of the RWH of long interest rates for low values of n, assume a deterministic system so that $\mu_{t+i} = 0$ and ${}_t\gamma_{t+i} = i\gamma$. Further, assume that there is a shift in fundamental factors – with the shift factor β already known in period t 1 – such that:

$$r_{t+1} = \cdots = r_{t+n} = r_t + \beta \tag{39}$$

where β is the amount of the shift. Given these assumptions, (39) and (35), for $\underline{n}$ 1, imply that the evolution of the long rate on two-period bonds would be given by:

$$\begin{aligned} {}_tR_{t+1} - {}_{t-1}R_t &= (1/2)(r_t + \beta - r_t) \\ &= (1/2)\beta \qquad (40) \end{aligned}$$

We also have:

$$ {}_tR_{t+2} - {}_{t-1}R_{t+1} = (1/3)\beta \qquad (41)$$

and so on to:

$$ {}_tR_{t+n} - {}_{t-1}R_{t+n-1} = (1/(n+1))\beta \qquad (42)$$

where the right-hand side goes to zero only as $n \to \infty$, so that either the RWH must be confined to very large values of n or we must assume that there is no change ($\underline{\beta}$ 0) in the fundamental or systematic determinants of the long rates.

Pesando (1978) tested the random walk hypothesis – with the assumption of a time-invariant liquidity premium term – in the form of the difference between the rationally expected long yield ($E_{t-1}\ {}_tR_{t+n}|I_{t-1}$), based on information in $t - 1$, and the long yield ${}_{t|1}R_{t\ n-1}$. His dependent variable, therefore, was $[(E_{t-1}\ {}_tR_{t\ n+}I_{t-1})\ \ {}_{t\ 1}R_{t\ n\ 1}]$, which was implied by the joint hypothesis to be random and uncorrelated with information available at the beginning of the period. His regressions of this variable were done against a number of variables such as investment and saving, government deficit, deficit on the current account of the balance of payments,[7] real monetary base and real GNP, and the current change in the monetary base.[8] Pesando's tests on Canadian ten-year bond yields for the periods 1961:1 to 1971:2 and 1961:1 to 1976:4 showed insignificant and/or incorrectly signed coefficients for these variables, so that he could not reject the hypothesis that the current change in the long-term bond yield is a random variable and follows a martingale sequence. Pesando's tests of the models used by Sargent (1969), Echols and Elliot (1976) and Feldstein and Eckstein (1970) for his Canadian data set led to their rejection. These results also rejected the notion that the long yield includes a cyclical term premium determined by these variables, thereby lending support to the hypothesis of a time-invariant premium. Further, his tests rejected the autoregressive procedure for modeling the expected inflation rate. Pesando also rejected the Modigliani and Sutch preferred habitat model since this model requires that the liquidity premium is not time invariant.

The alternative assumption to the above random walk hypothesis is that the changes in long yields depend on economic variables.

[7] These variables were also included in the Echols and Elliot (1976) study.

[8] These variables are part of the liquidity preference approach and were also included in the Feldstein and Eckstein (1970) study discussed in the preceding chapter.

Information content of the term structure for the expected rates of inflation

Fisher's relationship between the nominal and real forward interest rates in efficient markets can be specified as:

$$ {}_t r^{\mathrm{f}}{}_{t+i} = {}_t r^{\mathrm{re}}{}_{t+i} + {}_t \pi^{\mathrm{e}}{}_{t+i} \qquad (43)$$

where:

${}_t r^{\mathrm{f}}{}_{t+i}$ = forward market (nominal) rate of interest in period $t + i$

${}_t r^{\mathrm{re}}{}_{t+i}$ = expected real rate of interest in period $t + i$

${}_t \pi^{\mathrm{e}}{}_{t+i}$ = expected rate of inflation in period $t + i$.

Restate (43) as:

$$ {}_t \pi^{\mathrm{e}}{}_{t+i} = {}_t r^{\mathrm{f}}{}_{t+i} - {}_t r^{\mathrm{re}}{}_{t+i} \qquad (44)$$

Now use the REH to specify the following relationships:

$$ {}_t r^{\mathrm{re}}{}_{t+i} = r^{\mathrm{r}}{}_{t+i} - \eta_{t+i} \qquad (45)$$

$$ {}_t \pi^{\mathrm{e}}{}_{t+i} = E_t \pi_{t+i} \qquad (46)$$

Equation (44) can now be rewritten as:

$$ E_{tt}\pi_{t+i} = {}_t r^{\mathrm{f}}{}_{t+i} - {}_t r^{\mathrm{r}}{}_{t+i} + \eta_{t+i} \qquad (47)$$

which can be restated as:

$$ E_{tt}\pi_{T\ +i} = {}_t r^{\mathrm{f}}{}_{t+i} - E_{\ tt} r^{\mathrm{r}}{}_{\ +i} \qquad (48)$$

$$t$$

Equation (48) uses the information on the nominal forward rates, which are known, and the rationally expected value of the *real* interest rates – assuming the latter to be known or already estimated – to derive the expected inflation rates and their term structure over future periods.

Common hypotheses on the real rate of interest

Equations (47) and (48) require data on or estimates of the future real rate $r^{\mathrm{r}}{}_{t\,i}$. The range of choices here is usually as follows.

$+$

(i) Market data is available on it; e.g. if there is an adequate variety of inflation-indexed bonds in the economy.

(ii) Market data on the future real rate is not available but it can be reasonably assumed to be constant, or that changes in it are very small relative to changes in the inflation rate.

(iii) The assumption of its constancy is not plausible but it is a function of a small number of determinants and this function can be estimated. For example, the loanable funds or the liquidity preference theory, as discussed in Chapter 19, or the more general

IS–LM approach can be used to specify the determinants of this rate. Thus, the IS–LM approach implies that the real money supply and the real fiscal deficit are among the short-run determinants of the real interest rate. The appropriate function for it can be specified and estimated, and the estimated values then substituted in (47) or (48). Among the studies that use this approach to the real rate are those by Sargent (1969), Echols and Elliot (1976), Feldstein and Eckstein (1970) and Pesando (1978). Alternatively, the real interest rate may be assumed to be set by the central bank under a specified Taylor rule.

For an example of (ii), Walsh (2003, p. 498) reports the findings of a 1996 study by Buttiglione, Del Giovane and Tristani. This study examined the impact of monetary policy on long-term interest rates under the assumption that monetary policy does not affect real rates far in the future, so that such future real rates can be taken to be constant. Therefore, the change in long-term interest rates between future periods can be used to estimate the expected inflation rates. Walsh reported the findings of the Buttiglione *et al.* (1996) study as follows: for countries with low average inflation, a contractionary monetary policy raised short rates but lowered forward ones; whereas, for countries with high average inflation, the rise in short rates did not necessarily yield lower forward rates. Walsh points out that the finding for countries with low average inflation is consistent with the hypothesis that, in these countries, the restrictive monetary policy was viewed as a credible policy to lower inflation. Hence, the impact of monetary policy on long interest rates depends not only on the policy pursued but also on the evaluation by the public of its impact on the inflation rate.

An illustration of the empirical results

Mishkin (1990) examined the rates of inflation implied by the yields on one- to five-year bonds in the United States. His data used was monthly from 1953 to 1987. His estimation equation was:

$$ {}_t\pi_{t+i} - {}_t\pi_t = \alpha_{t+i} + \beta_{t+i}({}_tR_{t+i} - {}_tR_t) + \mu_{t+i} \qquad (49) $$

where the difference ("inflation spread") in the average annual inflation rate over t i years was regressed on the spread between the corresponding nominal long rates. Mishkin argued that a rejection of β_{t+i} 0 implies that the term structure contains information on the inflation spread, while a rejection of $\beta_{t\ \ i}$ 1 implies that spread in the real rates is not constant over time, in which case the nominal spread in the interest rates does not have any information on the inflation spread, so that nominal spread would provide information on the real spread. His estimates showed that, for the longer maturities, the spread in nominal interest rates contains substantial information about the inflation spread but little about the real interest rate spread. Contrary to the findings of many other studies, the converse was found for short maturities of six months or less. For these, the term structure did not contain any information on the future change in inflation but did imply a significant amount about the term structure of *real* rates of interest.

Barr and Campbell (1997) provide an example of the derivation of expected future increases in inflation from data on long rates. For the UK, they use the yield on indexed (i.e. with the nominal interest rising to compensate for the inflation rate) and nominal (un-indexed) government bonds to estimate the expected inflation rates. Their finding is that about 80 percent of the changes in the long-term nominal interest rates reflect expected long-term inflation.

Note that the spread between the interest rates on short-term and long-term bonds and those on commercial paper and Treasury bills can have predictive power, not only for expected inflation but also for future output changes, as reported, for example, by Stock and Watson (1989) and Friedman and Kuttner (1992).