# "Quantity Theory of Money"

by Milton Friedman

In *The New Palgrave: A Dictionary of Economics*, edited by John Eatwell, Murray Milgate, and
Peter Newman, vol. 4, pp. 3-20. New York: Stockton Press; and London: Macmillan, 1987.

> *Lowness of interest is generally ascribed to plenty of money. But … augmentation [in the quantity of money] has no other effect than to heighten the price of labour and commodities … In the progress toward these changes, the augmentation may have some influence, by exciting industry, but after the prices are settled … it has no manner of influence.*
>
> *[T]hough 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…. In my opinion, it is only in this interval of intermediate situation, between the acquisition of money and rise of prices, that the increasing quantity of gold and silver is favourable to industry…. [W]e may conclude that it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in greater or less quantity. The good policy of the magistrate consists only in keeping it, if possible, still increasing … (David Hume, 1752).*

In this survey, we shall first present a formal statement of the quantity theory, then consider the Keynesian challenge to the quantity theory, recent developments, and some empirical evidence. We shall conclude with a discussion of policy implications, giving special attention to the likely implications of the worldwide fiat money standard that has prevailed since 1971.

## 1. The Formal Theory

(a) NOMINAL VERSUS REAL QUANTITY OF MONEY. Implicit in the quotation from Hume, and central to all later versions of the quantity theory, is a distinction between the *nominal* quantity of money and the *real* quantity of money. The nominal quantity of money is the quantity expressed in whatever units are used to designate money – talents, shekels, pounds, francs, lira, drachmas, dollars, and so on. The real quantity of money is the quantity expressed in terms of the volume of goods and services the money will purchase.

There is no unique way to express either the nominal or the real quantity of money. With respect to the nominal quantity of money, the issue is what assets to include – whether only currency and coins, or also claims on financial institutions; and, if such claims are included, which ones should be, only deposits transferable by cheque, or also other categories of claims which in practice are close substitutes for deposits transferable by cheque. More recently, economists have been experimenting with the theoretically attractive idea of defining money not as the simple sum of various categories of claims but as a weighted aggregate of such claims, the weights being determined by one or another concept of the "moneyness" of the various claims.

Despite continual controversy over the definition of "money," and the lack of unanimity about relevant theoretical criteria, in practice, monetary economists have generally displayed wide agreement about the most useful counterpart, or set of counterparts, to the concept of "money" at particular times and places (Friedman and Schwartz, 1970, pp. 89–197; Barnett, Offenbacher and Spindt, 1984; Spindt, 1985).

The real quantity of money obviously depends on the particular definition chosen for the nominal quantity. In addition, for each such definition, it can vary according to the set of goods and services in terms of which it is expressed. One way to calculate the real quantity of money is by dividing the nominal quantity of money by a price index. The real quantity is then expressed in terms of the standard basket whose components are used as weights in computing the price index – generally, the basket purchased by some representative group in a base year.

A different way to express the real quantity of money is in terms of the time duration of the flow of goods and services the money could purchase. For a household, for example, the real quantity of money can be expressed in terms of the number of weeks of the household’s average level of consumption its money balances could finance or, alternatively, in terms of the number of weeks of its average income to which its money balances are equal. For a business enterprise, the real quantity of money it holds can be expressed in terms of the number of weeks of its average purchases, or of its average sales, or of its average expenditures on final productive services (net value added) to which its money balances are equal. For the community as a whole, the real quantity of money can be expressed in terms of the number of weeks of aggregate transactions of the community, or aggregate net output of the community, to which its money balances are equal.

The reciprocal of any of this latter class of measures of the real quantity of money is a velocity of circulation for the corresponding unit or group of units. For example, the ratio of the annual transactions of the community to its stock of money is the "transactions velocity of circulation of money," since it gives the number of times the stock of money would have to "turn over" in a year to accomplish all transactions. Similarly, the ratio of annual income to the stock of money is termed "income velocity." In every case, the real quantity of money is calculated at the set of prices prevailing at the date to which the calculation refers. These prices are the bridge between the nominal and the real quantity of money.

The quantity theory of money takes for granted, first, that the real quantity rather than the nominal quantity of money is what ultimately matters to holders of money and, second, that in any given circumstances people wish to hold a fairly definite real quantity of money. Starting from a situation in which the nominal quantity that people hold at a particular moment of time happens to correspond at current prices to the real quantity that they wish to hold, suppose that the quantity of money unexpectedly increases so that individuals have larger cash balances than they wish to hold. They will then seek to dispose of what they regard as their excess money balances by paying out a larger sum for the purchase of securities, goods, and services, for the repayment of debts, and as gifts, than they are receiving from the corresponding sources. However, they cannot as a group succeed. One man’s spending is another man’s receipts. One man can reduce his nominal money balances only by persuading someone else to increase his. The community as a whole cannot in general spend more than it receives; it is playing a game of musical chairs.

The attempt to dispose of excess balances will nonetheless have important effects. If prices and incomes are free to change, the attempt to spend more will raise total spending and receipts, expressed in nominal units, which will lead to a bidding up of prices and perhaps also to an increase in output. If prices are fixed by custom or by government edict, the attempt to spend more will either be matched by an increase in goods and services or produce "shortages" and "queues." These in turn will raise the effective price and are likely sooner or later to force changes in customary or official prices.

The initial excess of nominal balances will therefore tend to be eliminated, even though there is no change in the nominal quantity of money, by either a reduction in the real quantity available to hold through price rises or an increase in the real quantity desired through output increases. And conversely for an initial deficiency of nominal balances.

Changes in prices and nominal income can be produced either by changes in the real balances that people wish to hold or by changes in the nominal balances available for them to hold. Indeed, it is a tautology, summarized in the famous quantity equations, that all changes in nominal income can be attributed to one or the other – just as a change in the price of any good can always be attributed to a change in either demand or supply. The quantity theory is not, however, this tautology. On an analytical level, it has long been an analysis of the factors determining the quantity of money that the community wishes to hold; on an empirical level, it has increasingly become the generalization that changes in desired real balances (in the demand for money) tend to proceed slowly and gradually or to be the result of events set in train by prior changes in supply, whereas, in contrast, substantial changes in the supply of nominal balances can and frequently do occur independently of any changes in demand. The conclusion is that substantial changes in prices or nominal income are almost always the result of changes in the nominal supply of money.

(b) QUANTITY EQUATIONS. Attempts to formulate mathematically the relations just presented verbally date back several centuries (Humphrey, 1984). They consist of creating identities equating a flow of money payments to a flow of exchanges of goods or services. The resulting quantity equations have proved a useful analytical device and have taken different forms as quantity theorists have stressed different variables.

*The transactions form of the quantity equation*. The most famous version of the quantity equation is doubtless the transactions version formulated by Simon Newcomb (1885) and popularized by Irving Fisher (1911):

$$MV = PT, \qquad (1)$$

or

$$MV + M'V' = PT. \qquad (2)$$

In this version the elementary event is a transaction – an exchange in which one economic actor transfers goods or services or securities to another actor and receives a transfer of money in return. The right-hand side of the equations corresponds to the transfer of goods, services, or securities; the left-hand side, to the matching transfer of money.

Each transfer of goods, services or securities is regarded as the product of a price and quantity; wage per week times number of weeks, price of a good times number of units of the good, dividend per share times number of shares, price per share times number of shares, and so on. The right-hand side of equations (1) and (2) is the aggregate of such payments during some interval, with $P$ a suitably chosen *average* of the prices and $T$ a suitably chosen *aggregate* of the quantities during that interval, so that $PT$ is the total nominal value of the payments during the interval in question. The units of $P$ are dollars (or other monetary unit) per unit of quantity; the units of $T$ are number of unit quantities per period of time. We can convert the equation from an expression applying to an *interval* of time to one applying to a *point* in time by the usual limiting process of letting the interval for which we aggregate payments approach zero, and expressing $T$ not as an aggregate but as a rate of flow. The magnitude $T$ then has the dimension of quantity per unit time; the product of $P$ and $T$, of dollars (or other monetary unit) per unit time.

$T$ is clearly a rather special index of quantities: it includes service flows (man-hours, dwelling-years, kilowatt-hours) and also physical capital items yielding such flows (houses, electric-generating plants) and securities representing both physical capital items and such intangible capital items as "goodwill." Since each capital item or security is treated as if it disappeared from economic circulation once it is transferred, any such item that is transferred more than once in the period in question is implicitly weighted by the number of times it enters into transactions (its "velocity of circulation," in strict analogy with the "velocity of circulation" of money). Similarly, $P$ is a rather special price index.

The monetary transfer analysed on the left-hand side of equations (1) and (2) is treated very differently. The money that changes hands is treated as retaining its identity, and all money, whether used in transactions during the time interval in question or not, is explicitly accounted for. Money is treated as a stock, not as a flow or a mixture of a flow and a stock. For a single transaction, the breakdown into $M$ and $V$ is trivial: the cash that is transferred is turned over once, or $V$=1. For all transactions during an interval of time, we can, in principle, classify the existing stock of monetary units according as each monetary unit entered into 0, 1, 2, … transactions – that is, according as the monetary unit "turned over" 0, 1, 2, … times. The weighted average of these numbers of turnover, weighted by the number of dollars that turned over that number of times, is the conceptual equivalent of $V$. The dimensions of $M$ are dollars (or other monetary unit); of $V$, number of turnovers per unit time; so, of the product, dollars per unit time.

Equation (2) differs from equation (1) by dividing payments into two categories: those effected by the transfer of hand-to-hand currency (including coin) and those effected by the transfer of deposits. In equation (2) $M$ stands for the volume of currency and $V$ for the velocity of currency, $M'$ for the volume of deposits, and $V'$ for the velocity of deposits.

One reason for the emphasis on this particular division was the persistent dispute about whether the term *money* should include only currency or deposits as well. Another reason was the direct availability of data on $M'V'$ from bank records of clearings or of debits to deposit accounts. These data make it possible to calculate $V'$ in a way that is not possible for $V$.

Equations (1) and (2), like the other quantity equations we shall discuss, are intended to be identities – a special application of double-entry bookkeeping, with each transaction simultaneously recorded on both sides of the equation. However, as with the national income identities with which we are all familiar, when the two sides, or the separate elements on the two sides, are estimated from independent sources of data, many differences between them emerge. This statistical defect has been less obvious for the quantity equations than for the national income identities – with their standard entry "statistical discrepancy" – because of the difficulty of calculating $V$ directly. As a result, $V$ in equation (1) and $V$ and $V'$ in equation (2) have generally been calculated as the numbers having the property that they render the equations correct. These calculated numbers therefore embody the whole of the counterpart to the "statistical discrepancy."

Just as the left-hand side of equation (1) can be divided into several components, as in equation (2), so also can the right-hand side. The emphasis on transactions reflected in this version of the quantity equation suggests dividing total transactions into categories of payments for which payment periods or practices differ: for example, into capital transactions, purchases of final goods and services, purchases of intermediate goods, and payments for the use of resources, perhaps separated into wage and salary payments and other payments. The observed value of $V$ might well depend on the distribution of total payments among categories. Alternatively, if the quantity equation is interpreted not as an identity but as a functional relation expressing desired velocity as a function of other variables, the distribution of payments may well be an important set of variables.

*The income form of the quantity equation*. Despite the large amount of empirical work done on the transactions equations, notably by Irving Fisher (1911, pp. 280–318; 1919, pp. 407–9) and Carl Snyder (1934, pp. 278–91), the ambiguities of the concepts of "transactions" and the "general price level" – particularly those arising from the mixture of current and capital transactions – have never been satisfactorily resolved. More recently, national or social accounting has stressed income transactions rather than gross transactions and has explicitly if not wholly satisfactorily dealt with the conceptual and statistical problems involved in distinguishing between changes in prices and changes in quantities. As a result, since at least the work of James Angell (1936), monetary economists have tended to express the quantity equation in terms of income transactions rather than gross transactions. Let $Y$ = nominal income, $P$ = the price index implicit in estimating national income at constant prices, $N$ = the number of persons in the population, $y$ = per capita national income in constant prices, and $y' = Ny$ = national income at constant prices, so that

$$Y = PNy = Py'. \qquad (3)$$

Let $M$ represent, as before, the stock of money; but define $V$ as the average number of times per unit time that the money stock is used in making *income* transactions (that is, payment for final productive services or, alternatively, for final goods and services) rather than all transactions. We can then write the quantity equation in income form as

$$MV = PNy = Py'. \qquad (4)$$

or, if we desire to distinguish currency from deposit transactions, as

$$MV + M'V' = PNy. \qquad (5)$$

Although the symbols $P$,$V$, and $V'$ are used both in equations (4) and (5) and in equations (1) and (2), they stand for different concepts in each pair of equations. (In practice, gross national product often replaces national income in calculating velocity even though the logic underlying the equation calls for national income. The reason is the widespread belief that estimates of GNP are subject to less statistical error than estimates of national income.)

In the transactions version of the quantity equation, each intermediate transaction – that is, purchase by one enterprise from another – is included at the total value of the transaction, so that the value of wheat, for example, is included once when it is sold by the farmer to the mill, a second time when the mill sells flour to the baker, a third time when the baker sells bread to the grocer, a fourth time when the grocer sells bread to the consumer. In the income version, only the net value added by each of these transactions is included. To put it differently, in the transactions version, the elementary event is an isolated exchange of a physical item for money – an actual, clearly observable event. In the income version, the elementary event is a hypothetical event that can be inferred but is not directly observable. It is a complete series of transactions involving the exchange of productive services for final goods, via a sequence of money payments, with all the intermediate transactions in this income circuit netted out. The total value of all transactions is therefore a multiple of the value of income transactions only.

For a given flow of productive services or, alternatively, of final products (two of the multiple faces of income), the volume of transactions will be affected by vertical integration or disintegration of enterprises, which reduces or increases the number of transactions involved in a single income circuit, and by technological changes that lengthen or shorten the process of transforming productive services into final products. The volume of income will not be thus affected.

Similarly, the transactions version includes the purchase of an existing asset – a house or a piece of land or a share of equity stock – precisely on a par with an intermediate or final transaction. The income version excludes such transactions completely.

Are these differences an advantage or disadvantage of the income version? That clearly depends on what it is that determines the amount of money people want to hold. Do changes of the kind considered in the preceding paragraphs, changes that alter the ratio of intermediate and capital transactions to income, also alter in the same direction and by the same proportion the amount of money people want to hold? Or do they tend to leave this amount unaltered? Or do they have a more complex effect?

The transactions and income versions of the quantity theory involve very different conceptions of the role of money. For the transactions version, the most important thing about money is that it is transferred. For the income version, the most important thing is that it is held. This difference is even more obvious from the Cambridge cash-balance version of the quantity equation (Pigou, 1917). Indeed, the income version can perhaps best be regarded as a way station between the Fisher and the Cambridge version.

*Cambridge cash-balance approach*. The essential feature of a money economy is that an individual who has something to exchange need not seek out the double coincidence – someone who both wants what he has and offers in exchange what he wants. He need only find someone who wants what he has, sell it to him for general purchasing power, and then find someone who has what he wants and buy it with general purchasing power.

For the act of purchase to be separated from the act of sale, there must be something that everybody will accept in exchange as "general purchasing power" – this aspect of money is emphasized in the transactions approach. But also there must be something that can serve as a temporary abode of purchasing power in the interim between sale and purchase. This aspect of money is emphasized in the cash-balance approach.

How much money will people or enterprises want to hold on the average as a temporary abode of purchasing power? As a first approximation, it has generally been supposed that the amount bears some relation to income, on the assumption that income affects the volume of potential purchases for which the individual or enterprise wishes to hold cash balances. We can therefore write

$$M = kPNy = kPy'. \qquad (6)$$

where *M*, *N*, *P*, *y*, and *y′* are defined as in equation (4) and *k* is the ratio of money stock to income – either the observed ratio so calculated as to make equation (6) an identity or the "desired" ratio so that *M* is the "desired" amount of money, which need not be equal to the actual amount. In either case, *k* is numerically equal to the reciprocal of the *V* in equation (4), the *V* being interpreted in one case as measured velocity and in the other as desired velocity.

Although equation (6) is simply a mathematical transformation of equation (4), it brings out sharply the difference between the aspect of money stressed by the transactions approach and that stressed by the cash-balance approach. This difference makes different definitions of money seem natural and leads to placing emphasis on different variables and analytical techniques.

The transactions approach makes it natural to define money in terms of whatever serves as the medium of exchange in discharging obligations. The cash-balance approach makes it seem entirely appropriate to include in addition such temporary abodes of purchasing power as demand and time deposits not transferable by check, although it clearly does not require their inclusion (Friedman and Schwartz, 1970, ch. 3).

Similarly, the transactions approach leads to emphasis on the mechanical aspect of the payments process: payments practices, financial and economic arrangements for effecting transactions, the speed of communication and transportation, and so on (Baumol, 1952; Tobin, 1956; Miller and Orr, 1966, 1968). The cash-balance approach, on the other hand, leads to emphasis on variables affecting the usefulness of money as an asset: the costs and returns from holding money instead of other assets, the uncertainty of the future, and so on (Friedman, 1956; Tobin, 1958).

Of course, neither approach enforces the exclusion of the variables stressed by the other. Portfolio considerations enter into the costs of effecting transactions and hence affect the most efficient payment arrangements; mechanical considerations enter into the returns from holding cash and hence affect the usefulness of cash in a portfolio.

Finally, with regard to analytical techniques, the cash-balance approach fits in much more readily with the general Marshallian demand-supply apparatus than does the transactions approach. Equation (6) can be regarded as a demand function for money, with $P$, $N$, and $y$ on the right-hand side being three of the variables on which the quantity of money demanded depends and $k$ symbolizing all the other variables, so that $k$ is to be regarded not as a numerical constant but as itself a function of still other variables. For completion, the analysis requires another equation showing the supply of money as a function of these and other variables. The price level or the level of nominal income is then the resultant of the interaction of the demand and supply functions.

*Levels versus rates of change*. The several versions of the quantity equations have all been stated in terms of the levels of the variables involved. For the analysis of monetary change it is often more useful to express them in terms of rates of change. For example, take the logarithm of both sides of equation (4) and differentiate with respect to time. The result is

$$\frac{1}{M}\frac{\mathrm{d}M}{\mathrm{d}t}+\frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}t}=\frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}t}+\frac{1}{y'}\frac{\mathrm{d}y'}{\mathrm{d}t} \qquad (7)$$

or, in simpler notation,

$$g_M + g_V = g_p + g_{y'} = g_{Y'}, \qquad (8)$$

where $g$ stands for the percentage rate of change (continuously compounded) of the variable denoted by its subscript. The same equation is implied by equation (6), with $g_V$ replaced by $-g_k$.

The rate of change equations serve two very different purposes. First, they make explicit an important difference between a once-for-all change in the level of the quantity of money and a change in the rate of change of the quantity of money. The former is equivalent simply to a change of units – to substituting cents for dollars or pence for pounds – and hence, as is implicit in equations (4) and (6), would not be presumed to have any effect on real quantities, on neither $V$ (nor $k$) nor $y'$, but simply an offsetting effect on the price level, $P$. A change in the rate of change of money is a very different thing. It will tend, according to equations (7) and (8), to be accompanied by a change in the rate of inflation ($g_P$) which, as pointed out in section *d* below, affects the cost of holding money, and hence the desired real quantity of money. Such a change will therefore affect real quantities, $V$ and $g_V$, $y'$ and $g_y'$, as well as nominal and real interest rates.

The second purpose served by the rate of change equations is to make explicit the role of time, and thereby to facilitate the study of the effect of monetary change on the temporal pattern of response of the several variables involved. In recent decades, economists have devoted increasing attention to the short-term pattern of economic change, which has enhanced the importance of the rate of change versions of the quantity equations.

(c) THE SUPPLY OF MONEY. The quantity theory in its cash-balance version suggests organizing an analysis of monetary phenomena in terms of (1) the conditions determining supply (this section); (2) the conditions determining demand (section d below); and (3) the reconciliation of demand with supply (section e below).

The factors determining the nominal supply of money available to be held depend critically on the monetary system. For systems like those that have prevailed in most major countries during the past two centuries, they can usefully be analysed under three main headings termed the proximate determinants of the quantity of money: (1) the amount of high-powered money – specie plus notes or deposit liabilities issued by the monetary authorities and used either as currency or as reserves by banks; (2) the ratio of bank deposits to bank holdings of high-powered money; and (3) the ratio of the public's deposits to its currency holdings (Friedman and Schwartz, 1963b, pp. 776–98; Cagan, 1965; Burger, 1971; Black, 1975).

It is an identity that

$$M = H \cdot \frac{\frac{D}{R}\left(1+\frac{D}{C}\right)}{\frac{D}{R}+\frac{D}{C}}, \qquad (9)$$

where *H* = high-powered money; *D* = deposits; *R* = bank reserves; *C* = currency in the hands of the public so that (*D/R*) is the deposit-reserve ratio; and (*D/C*) is the deposit-currency ratio. The fraction on the right-hand side of (9), i.e., the ratio of *M* to *H*, is termed the money multiplier, often a convenient summary of the effect of the two deposit ratios. The determinants are called proximate because their values are in turn determined by much more basic variables. Moreover, the same labels can refer to very different contents.

High-powered money is the clearest example. Until some time in the 18th or 19th century, the exact date varying from country to country, it consisted only of specie or its equivalent: gold, or silver, or cowrie shells, or any of a wide variety of commodities. Thereafter, until 1971, with some significant if temporary exceptions, it consisted of a mixture of specie and of government notes or deposit liabilities. The government notes and liabilities generally were themselves promises to pay specified amounts of specie on demand, though this promise weakened after World War I, when many countries promised to pay either specie or foreign currency. During the Bretton Woods periods after World War II, only the USA was obligated to pay gold, and only to foreign monetary agencies, not to individuals or other non-governmental entities; other countries obligated themselves to pay dollars.

Since 1971, the situation has been radically different. In every major country, high-powered money consists solely of fiat money – pieces of paper issued by the government and inscribed with the legend "one dollar" or "one pound" and the message "legal tender for all debts public and private"; or book entries, labeled deposits, consisting of promises to pay such pieces of paper. Such a worldwide fiat (or irredeemable paper) standard has no precedent in history. The "gold" that central banks still record as an asset on their books is simply the grin of a Cheshire cat that has disappeared.

Under an international commodity standard, the total quantity of high-powered money in any one country – so long as it remains on the standard – is determined by the balance of payments. The division of high-powered money between physical specie and the fiduciary component of government-issued promises to pay is determined by the policies of the monetary authorities. For the world as a whole, the total quantity of high-powered money is determined both by the policies of the various monetary authorities and the physical conditions of supply of specie. The latter provide a physical anchor for the quantity of money and hence ultimately for the price level.

Under the current international fiat standard, the quantity of high-powered money is determined solely by the monetary authorities, consisting in most countries of a central bank plus the fiscal authorities. What happens to the quantity of high-powered money depends on their objectives, on the institutional and political arrangements under which they operate, and the operating procedures they adopt. These are likely to vary considerably from country to country. Some countries (e.g., Hong Kong, Panama) have chosen to link their currencies rigidly to some other currency by pegging the exchange rate. For them, the amount of high-powered money is determined in the same way as under an international commodity standard – by the balance of payments.

The current system is so new that it must be regarded as in a state of transition. Some substitute is almost sure to emerge to replace the supply of specie as a long-term anchor for the price level, but it is not yet clear what that substitute will be (see section 5 below).

*The deposit-reserve ratio* is determined by the banking system subject to any requirements that are imposed by law or the monetary authorities. In addition to any such requirements, it depends on such factors as the risk of calls for conversion of bank deposits to high-powered money; the cost of acquiring additional high-powered money in case of need; and the returns from loans and investments, that is, the structure of interest rates.

*The deposit-currency ratio* is determined by the public. It depends on the relative usefulness to holders of money of deposits and currency and the relative cost of holding the one or the other. The relative cost in turn depends on the rates of interest received on deposits, which may be subject to controls imposed by law or the monetary authorities.

These factors determine the *nominal*, but not the *real*, quantity of money. The real quantity of money is determined by the interaction between the *nominal* quantity supplied and the *real* quantity demanded. In the process, changes in demand for real balances have feedback effects on the variables determining the nominal quantity supplied, and changes in nominal supply have feedback effects on the variables determining the real quantity demanded. Quantity theorists have generally concluded that these feedback effects are relatively minor, so that the *nominal* supply can generally be regarded as determined by a set of variables distinct from those that affect the *real* quantity demanded. In this sense, the nominal quantity can be regarded as determined primarily by supply, the real quantity, primarily by demand.

Instead of expressing the nominal supply in terms of the identity (9), it can also be expressed as a function of the variables that are regarded as affecting H, D/R, and D/C, such as the rate of inflation, interest rates, nominal income, the extent of uncertainty, perhaps also the variables that are regarded as determining the decisions of the monetary authorities. Such a supply function is frequently written as

$$M^S = h(R, Y, \ldots), \qquad (10)$$

where *R* is an interest rate or set of interest rates, *Y* is nominal income, and the dots stand for other variables that are regarded as relevant.

(d) THE DEMAND FOR MONEY. The cash-balance version of the quantity theory, by stressing the role of money as an asset, suggests treating the demand for money as part of capital or wealth theory, concerned with the composition of the balance sheet or portfolio of assets.

From this point of view, it is important to distinguish between ultimate wealth holders, to whom money is one form in which they choose to hold their wealth, and enterprises, to whom money is a producer's good like machinery or inventories (Friedman, 1956; Laidler, 1985; Friedman and Schwartz, 1982).

*Demand by ultimate wealth holders*. For ultimate wealth holders the demand for money, in real terms, may be expected to be a function primarily of the following variables:

1. *Total wealth*. This is the analogue of the budget constraint in the usual theory of consumer choice. It is the total that must be divided among various forms of assets. In practice, estimates of total wealth are seldom available. Instead, income may serve as an index of wealth. However, it should be recognized that income as measured by statisticians may be a defective index of wealth because it is subject to erratic year-to-year fluctuations, and a longer-term concept, like the concept of permanent income developed in connection with the theory of consumption, may be more useful (Friedman, 1957, 1959).

The emphasis on income as a surrogate for wealth, rather than as a measure of the "work" to be done by money, is perhaps the basic conceptual difference between the more recent analyses of the demand for money and the earlier versions of the quantity theory.

2. *The division of wealth between human and non-human forms*. The major asset of most wealth holders is personal earning capacity. However, the conversion of human into non-human wealth or the reverse is subject to narrow limits because of institutional constraints. It can be done by using current earnings to purchase non-human wealth or by using non-human wealth to finance the acquisition of skills, but not by purchase or sale of human wealth and to only a limited extent by borrowing on the collateral of earning power. Hence, the fraction of total wealth that is in the form of non-human wealth may be an additional important variable.

3. *The expected rates of return on money and other assets*. These rates of return are the counterparts to the prices of a commodity and its substitutes and complements in the usual theory of consumer demand. The nominal rate of return on money may be zero, as it generally is on currency, or negative, as it sometimes is on demand deposits subject to net service charges, or positive, as it sometimes is on demand deposits on which interest is paid and generally is on time deposits. The nominal rate of return on other assets consists of two parts: first, any currently paid yield, such as interest on bonds, dividends on equities, or cost, such as storage costs on physical assets, and, second, a change in the nominal price of the asset. The second part is especially important under conditions of inflation or deflation.

4. *Other variables determining the utility attached to the services rendered by money relative to those rendered by other assets – in Keynesian terminology, determining the value attached to liquidity proper*. One such variable may be one already considered – namely, real wealth or

income, since the services rendered by money may, in principle, be regarded by wealth holders as a "necessity," like bread, the consumption of which increases less than in proportion to any increase in income, or as a "luxury," like recreation, the consumption of which increases more than in proportion.

Another variable that is important empirically is the degree of economic stability expected to prevail, since instability enhances the value wealth-holders attach to liquidity. This variable has proved difficult to express quantitatively although qualitative information often indicates the direction of change. For example, the outbreak of war clearly produces expectations of greater instability. That is one reason why a notable increase in real balances – that is, a notable decline in velocity – often accompanies the outbreak of war. Such a decline in velocity produced an initial *decline* in sensitive prices at the outset of both World War I and World War II – not the rise that later inflation would have justified.

The rate of inflation enters under item 3 as a factor affecting the cost of holding various assets, particularly currency. The variability of inflation enters here, as a major factor affecting the usefulness of money balances. Empirically, variability of inflation tends to increase with the level of inflation, reinforcing the negative effect of higher inflation on the quantity of money demanded.

Still another relevant variable may be the volume of trading in existing capital goods by ultimate wealth holders. The higher the turnover of capital assets, the larger the fraction of total assets people may find it useful to hold as cash. This variable corresponds to the class of transactions omitted in going from the transactions version of the quantity equation to the income version.

We can express this analysis in terms of the following demand function for money for an individual wealth holder:

$$M^D = P \cdot f(y, w; R_M^*, R_B^*, R_E^*; u), \qquad (11)$$

where *M, P*, and *y* have the same meaning as in equation (6) except that they relate to a single wealth-holder (for whom *y = y′); w* is the fraction of wealth in non-human form (or, alternatively, the fraction of income derived from property); an asterisk denotes an expected value, so $R_M^*$ is the expected nominal rate of return on money; $R_B^*$ is the expected nominal rate of return on fixed-value securities, including expected changes in their prices; $R_E^*$ is the expected nominal rate of return on physical assets, including expected changes in their prices; and *u* is a portmanteau symbol standing for other variables affecting the utility attached to the services of money. Though the expected rate of inflation is not explicit in equation (11), it is implicit because it affects the expected nominal returns on the various classes of assets, and is sometimes used as a proxy for $R_E^*$. For some purposes it may be important to classify assets still more finely – for example, to distinguish currency from deposits, long-term from short-term fixed-value securities, risky from relatively safe equities, and one kind of physical assets from another.

Furthermore, the several rates of return are not independent. Arbitrage tends to eliminate differences among them that do not correspond to differences in perceived risk or other nonpecuniary characteristics of the assets, such as liquidity. In particular, as Irving Fisher pointed out in 1896, arbitrage between real and nominal assets introduces an allowance for anticipated inflation into the nominal interest rate (Fisher, 1896; Friedman, 1956).

The usual problems of aggregation arise in passing from equation (11) to a corresponding equation for the economy as a whole – in particular, from the possibility that the amount of money demanded may depend on the distribution among individuals of such variables as *y* and *w* and not merely on their aggregate or average value. If we neglect these distributional effects, equation (11) can be regarded as applying to the community as a whole, with *M* and *y* referring to per capita money holdings and per capita real income, respectively, and *w* to the fraction of aggregate wealth in non-human form.

Although the mathematical equation may be the same, its significance is very different for the individual wealth-holder and the community as a whole. For the individual, all the variables in the equation other than his own income and the disposition of his portfolio are outside his control. He takes them, as well as the structure of monetary institutions, as given, and adjusts his nominal balances accordingly. For the community as a whole, the situation is very different. In general, the nominal quantity of money available to be held is fixed and what adjusts are the variables on the right-hand side of the equation, including an implicit underlying variable, the structure of monetary institutions, which, in the longer run, at least, adjusts itself to the tastes and preferences of the holders of money. A dramatic example is provided by the restructuring of the financial system in the US in the 1970s and 1980s.

In practice, the major problems that arise in applying equation (11) are the precise definitions of *y* and *w*, the estimation of *expected* rates of return as contrasted with actual rates of return, and the quantitative specification of the variables designated by *u*.

*Demand for business enterprises*. Business enterprises are not subject to a constraint comparable to that imposed by the total wealth of the ultimate wealth-holder. They can determine the total amount of capital embodied in productive assets, including money, to maximize returns, since they can acquire additional capital through the capital market.

A similar variable defining the "scale" of the enterprise may, however, be relevant as an index of the productive value of different quantities of money to the enterprise. Lack of data has meant that much less empirical work has been done on the business demand for money than on an aggregate demand curve encompassing both ultimate wealth-holders and business enterprises. As a result, there are as yet only faint indications about the best variable to use: whether total transactions, net value added, net income, total capital in nonmoney form, of net worth.

The division of wealth between human and non-human form has no special relevance to business enterprises, since they are likely to buy the services of both forms on the market.

Rates of return on money and on alternative assets are, of course, highly relevant to business enterprises. These rates determine the net cost of holding money balances. However, the particular rates that are relevant may differ from those that are relevant for ultimate wealth-holders. For example, the rates banks charge on loans are of minor importance for wealth-holders yet may be extremely important for businesses, since bank loans may be a way in which they can acquire the capital embodied in money balances.

The counterpart for business enterprises of the variable *u* in equation (11) is the set of variables other than scale affecting the productivity of money balances. At least one subset of such variables – namely, expectations about economic stability and the variability of inflation – is likely to be common to business enterprises and ultimate wealth-holders.

With these interpretations of the variables, equation (11), with *w* excluded, can be regarded as symbolizing the business demand for money and, as it stands, symbolizing aggregate demand for money, although with even more serious qualifications about the ambiguities introduced by aggregation.

*Buffer stock effects*. In serving its basic function as a temporary abode of purchasing power, cash balances necessarily fluctuate, absorbing temporary discrepancies between the purchases and sales they mediate.

Though always recognized, this "buffer stock" role of money has seldom been explicitly modelled. Recently, more explicit attention has been paid to the buffer stock notion in an attempt to explain anomalies that have arisen in econometric estimates of the short-run demand for money (Judd and Scadding, 1982; Laidler, 1984; Knoester, 1984).

(e) THE RECONCILIATION OF DEMAND WITH SUPPLY. Multiply equation (11) by *N* to convert it from a per capita to an aggregate demand function, and equate it to equation (10), omitting for simplicity the asterisks designating expected values, and letting *R* stand for a vector of interest rates:

$$M^S = h(R, Y, \ldots) = P \cdot N \cdot f(y, w, R, g_P, u). \qquad (12)$$

The result is quantity equation (6) in an expanded form. In principle, a change in any of the underlying variables that produces a change in $M^s$ and disturbs a pre-existing equilibrium can produce offsetting changes in any of the other variables. In practice, as already noted earlier, the initial impact is likely to be on *y* and *R*, the ultimate impact predominantly on *P*.

A frequent criticism of the quantity theory is that its proponents do not specify the transmission mechanism between a change in $M^s$ and the offsetting changes in other variables, that they rely on a black box connecting the input – the nominal quantity of money – and the output – effects on prices and quantities.

This criticism is not justified insofar as it implies that the transmission mechanism for the quantity equation is fundamentally different from that for a demand–supply analysis of a particular product – shoes, or copper, or haircuts. In both cases the demand function for the community as a whole is the sum of demand functions for individual consumer or producer units, and the separate demand functions are determined by the tastes and opportunities of the units. In both cases, the supply function depends on production possibilities, institutional arrangements for organizing production, and the conditions of supply of resources. In both cases a shift in supply or in demand introduces a discrepancy between the amounts demanded and supplied *at the pre-existing price*. In both cases any discrepancy can be eliminated only by either a price change or some alternative rationing mechanism, explicit or implicit.

Two features of the demand–supply adjustment for money have concealed this parallelism. One is that demand-supply analysis for particular products typically deals with flows – number of pairs of shoes or number of haircuts per year – whereas the quantity equations deal with the stock of money at a point in time. In this respect the correct analogy is with the demand for, say, land, which, like money, derives its value from the flow of services it renders but has a purchase price and not merely a rental value. The second is the widespread tendency to confuse "money" and "credit," which has produced misunderstanding about the relevant price variable. The "price" of money is the quantity of goods and services that must be given up to acquire a unit of money – the inverse of the price level. This is the price that is analogous to the price of land or of copper or of haircuts. The "price" of money is not the interest rate, which is the "price" of credit. The interest rate connects stocks with flows – the rental value of land with the price of land, the value of the service flow from a unit of money with the price of money. Of course, the interest rate may affect the quantity of money demanded – just as it may affect the quantity of land demanded – but so may a host of other variables.

The interest rate has received special attention in monetary analysis because, without quite realizing it, fractional reserve banks have created part of the stock of money in the course of serving as an intermediary between borrowers and lenders. Hence changes in the quantity of money have frequently occurred through the credit markets, in the process producing important transitory effects on interest rates.

On a more sophisticated level, the criticism about the transmission mechanism applies equally to money and to other goods and services. In all cases it is desirable to go beyond equality of demand and supply as defining a stationary equilibrium position and examine the variables that affect the quantities demanded and supplied and the dynamic temporal process whereby actual or potential discrepancies are eliminated. Examination of the variables affecting demand and supply has been carried farther for money than for most other goods or services. But for both, there is as yet no satisfactory and widely accepted description, in precise quantifiable terms, of the dynamic temporal process of adjustment. Much research has been devoted to this question in recent decades; yet it remains a challenging subject for research. (For surveys of some of the literature, see Laidler, 1985; Judd and Scadding, 1982.)

(f) FIRST-ROUND EFFECTS. Another frequent criticism of the quantity equations is that they neglect any effect on the outcome of the source of change in the quantity of money. In Tobin’s words, the question is whether "the genesis of new money makes a difference," in particular, whether "an increase in the quantity of money has the same effect whether it is issued to purchase goods or to purchase bonds" (1974, p. 87).

Or, as John Stuart Mill put a very similar view in 1844, "The issues of a *Government* paper, even when not permanent, will raise prices; because Governments usually issue their paper in purchases for consumption. If issued to pay off a portion of the national debt, we believe they would have no effect" (1844, p. 589).

Tobin and Mill are right that the way the quantity of money is increased affects the outcome in some measure or other. If one group of individuals receives the money on the first round, they will likely use it for different purposes than another group of individuals. If the newly printed money is spent on the first round for goods and services, it adds directly at that point to the demand for such goods and services, whereas if it is spent on purchasing debt, or simply held temporarily as a buffer stock, it has no immediate effect on the demand for goods and services. Such effects come later as the initial recipients of the "new" money dispose of it. However, as the "new" money spreads through the economy, any first-round effects tend to be dissipated. The "new" money is merged with the old and is distributed in much the same way.

One way to characterize the Keynesian approach (see below) is that it gives almost exclusive importance to the first-round effect by putting primary emphasis on flows of spending rather than on stocks of assets. Similarly, one way to characterize the quantity-theory approach is to say that it gives almost no importance to first-round effects.

The empirical question is how important the first-round effects are compared with the ultimate effects. Theory cannot answer that question. The answer depends on how different are the reactions of the recipients of cash via alternative routes, on how rapidly a larger money stock is distributed through the economy, on how long it stays at each point in the economy, on how much the demand for money depends on the structure of government liabilities, and so on. Casual empiricism yields no decisive answer. Maybe the first-round effect is so strong that it dominates later effects; maybe it is highly transitory.

Despite repeated assertions by various authors that the first-round effect is significant, none, so far as I know, has presented any systematic empirical evidence to support that assertion. The apparently similar response of spending to changes in the quantity of money at widely separated dates in different countries and under diverse monetary systems establishes something of a presumption that the first-round effect is not highly significant. This presumption is also supported by several empirical studies designed to test the importance of the first-round effect (Cagan, 1972).

(g) THE INTERNATIONAL TRANSMISSION MECHANISM. From its very earliest days, the quantity theory was intimately connected with the analysis of the adjustment mechanism in international trade. A commodity standard, in which money is specie or its equivalent, was taken as the norm. Under such a standard, the supply of money in any one country is determined by the links between that country and other countries that use the same commodity as money. Under such a standard, the same theory explains links among money, prices, and nominal income in various parts of a single country — money, prices, and nominal income in Illinois and money, prices, and nominal income in the rest of the United States – and the corresponding links among various countries. The differences between interregional adjustment and international adjustment are empirical: greater mobility of people, goods, and capital among regions than among countries, and hence more rapid adjustment.

According to the specie-flow mechanism developed by Hume and elaborated by Henry Thornton, David Ricardo and their successors, "too" high a money stock in country A tends to make prices in A high relative to prices in the rest of the world, encouraging imports and discouraging exports. The resulting deficit in the balance of trade is financed by shipment of specie, which reduces the quantity of money in country A and increases it in the rest of the world, These changes in the quantity of money tend to lower prices in country A and raise them in the rest of the world, correcting the original disequilibrium. The process continues until price levels in all countries are at a level at which balances of payments are in equilibrium (which may be consistent with a continuing movement of specie, for example, from gold- or silver-producing countries to non-gold- or silver-producing countries, or between countries growing at different secular rates).

Another strand of the classical analysis has recently been revived under the title "the monetary theory of the balance of payments." The specie-flow mechanism implicitly assumes that prices adjust only in response to changes in the quantity of money produced by specie flows. However, if markets are efficient and transportation costs are neglected, there can be only a single price expressed in a common currency for goods traded internationally. Speculation tends to assure this result. Internally, competition between traded and nontraded goods tends to keep their relative price in line with relative costs. If these adjustments are rapid, "the law of one price" holds among countries. If the money stock is not distributed among countries in such a way as to be consistent with the equilibrium prices, excess demands and supplies of money will lead to specie flows. Domestic nominal demand in a country with "too" high a quantity of money will exceed the value of domestic output and the excess will be met by imports, producing a balance of payments deficit financed by the export of specie; and conversely in a country with too "low" a quantity of money. Specie flows are still the adjusting mechanism, but they are produced by differences between demand for output in nominal terms and the supply of output at world prices rather than by discrepancies in prices. Putative rather than actual price differences are the spur to adjustment. This description is highly oversimplified, primarily because it omits the important role assigned to short- and long-term capital flows by all theorists – those who stress the specie-flow mechanism and even more those who stress the single-price mechanism (Frenkel, 1976; Frenkel and Johnson, 1976).

In practice, few countries have had pure commodity standards. Most have had a mixture of commodity and fiduciary standards. Changes in the fiduciary component of the stock of money can replace specie flows as a means of adjusting the quantity of money.

The situation is still different for countries that do not share a unified currency, that is, a currency in which only the name assigned to a unit of currency differs among countries. Changes in the rates of exchange between national currencies then serve to keep prices in various countries in the appropriate relation when expressed in a common currency. Exchange rate adjustments replace specie flows or changes in the quantity of domestically created money. And exchange rate changes too may be produced by actual or putative price differences or by short- or long-term capital flows. Moreover, especially during the Bretton Woods period (1945–71), but more recently as well, governments have often tried to avoid changes in exchange rates by seeking adjustment through subsidies to exports, obstacles to imports, and direct controls over foreign exchange transactions. These measures involved either implicit or explicit multiple rate systems and were accompanied by government borrowing to finance balance-of payments deficits, or governmental lending to offset surpluses. They sometimes led to severe financial crises and major exchange rate adjustments – one reason the Bretton Woods system finally broke down in 1971. Since then, exchange rates have supposedly been free to float and to be determined in private markets. In practice, however, governments still intervene in an attempt to affect the exchange rates of their currencies, either directly by buying or selling their currency on the market, or indirectly, by adopting monetary or fiscal or trade policies designed to alter the market exchange rate. However, most governments no longer announce fixed parities for their currencies.

## 2 Keynesian Challenge to the Quantity Theory

The depression of the 1930s produced a wave of scepticism about the relevance and validity of the quantity theory of money. The central banks of the world – the Federal Reserve in the forefront – proclaimed that, despite the teachings of the quantity theory, "easy money" was proving to be ineffective in stemming the depression. They pointed to the low level of short-term interest rates as evidence of how "easy" monetary policy was. Their claims seemed credible not only because of the confusion between "lowness of interest" and "plenty of money" pointed out by Hume but also because of the absence of readily available evidence on what was happening to the quantity of money. Most observers at the time did not know, as we do now, that the Federal Reserve permitted the quantity of money in the United States to decline by one-third between 1929 and 1933, and hence that the accompanying contraction in economic activity and deflation of prices was entirely consistent with the quantity theory. Monetary policy was incredibly "tight" not "easy."

The scepticism about the quantity theory was further heightened by the publication of John Maynard Keynes’s *The General Theory of Employment, Interest and Money* (Keynes, 1936) which offered an alternative interpretation of economic fluctuations in general and the depression in particular. Keynes emphasized spending on investment and the stability of the consumption function rather than the stock of money and the stability of the demand function for money. He relegated the forces embodied in the quantity theory to a minor role, and treated fiscal rather than monetary policy as the chief instrument for influencing the course of events. Received wisdom both inside and outside the economics profession became "money does not matter."

Keynes did not deny the validity of the quantity equation, in any of its forms – after all, he had been a major contributor to the quantity theory (Keynes, 1923). What he did was something very different. He argued that the demand for money, which he termed the liquidity-preference function, had a special form such that *under conditions of underemployment* the $V$ in equation (4) and the $k$ in equation (6) would be highly unstable and would passively adapt to whatever changes independently occurred in money income or the stock of money. Under such conditions, these equations, though entirely valid, were largely useless for policy or prediction. Moreover, he regarded such conditions as prevailing much, if not most of the time.

That possibility rested on two other key propositions. First, that, contrary to the teachings of classical and neoclassical economists, the *long-run equilibrium* position of an economy need not be characterized by "full employment" of resources even if all prices are flexible. In his view, unemployment could be a deep-seated characteristic of an economy rather than simply a reflection of price and wage rigidity or transitory disturbances. This proposition has played an important role in promoting the acceptance of Keynesianism, especially by non-economists, even though, by now, it is widely accepted that, as a *theoretical* matter, the proposition is false. Keynes’s error consisted in neglecting the role of wealth in the consumption function. There is no fundamental "flaw in the price system" that makes persistent structural unemployment a possible or probable natural outcome of a fully operative market system (Haberler, 1941, pp. 242, 389, 403, 491–503; Pigou, 1947; Tobin, 1947; Patinkin, 1948; Johnson, 1961). The concept of "underemployment equilibrium" has been replaced by the concept of a "natural rate of unemployment" (see section 3 below).

Keynes's final key proposition was that, as an *empirical* matter, prices, especially wages, can be regarded as rigid – an institutional datum – for *short-run economic fluctuations*; in which case, the distinction between real and nominal magnitudes that is at the heart of the quantity theory is irrelevant for such fluctuations. This proposition, unlike the other two, did not conflict with the teachings of the quantity theory. Classical and neoclassical economists had long recognized that price and wage rigidity existed and contributed to unemployment during cyclical contractions, and to labour scarcity during cyclical booms. But to them, wage rigidity was a defect of the market; to Keynes, it was a rational response to the possibility of underemployment equilibrium (Keynes, 1936, pp. 269–71).

In his analysis of the demand for money (i.e., the form of equation (6) or (11)), Keynes treated the stock of money as if it were divided into two parts, one part, $M_1$ "held to satisfy the transactions- and precautionary-motives," the other, $M_2$, "held to satisfy the speculative-motive" (Keynes, 1936, p. 199). He regarded $M_1$, as a roughly constant fraction of income. He regarded the demand for $M_2$ as arising from "*uncertainty* as to the future course of the rate of interest" (Keynes, 1936, p. 168) and the amount demanded as depending on the relation between current rates of interest and the rates of interest expected to prevail in the future. Keynes, of course, recognized the existence of a whole complex of interest rates. However, for simplicity, he spoke in terms of "the rate of interest," usually meaning by that the rate on long-term securities that were fixed in nominal value and that involved minimal risks of default – for example, government bonds. In a "given state of expectations," the higher the current rate of interest, the lower would be the (real) amount of money that people would want to hold for speculative motives for two reasons: first, the greater would be the cost in terms of current earnings sacrificed by holding money instead of securities, and, second, the more likely it would be that interest rates would fall, and hence bond prices rise, and so the greater would be the cost in terms of capital gains sacrificed by holding money instead of securities.

To formalize Keynes's analysis in terms of the symbols we have used so far, we can write his demand (liquidity-preference) function as

$$M / P = M_1 / P + M_2 / P = k_1 y' + f(R - R^*, R^*) \qquad (13)$$

where $R$ is the current rate of interest, $R^*$ is the rate of interest expected to prevail, and $k_1$, the analogue to the inverse of the income velocity of circulation of money, is treated as determined by payment practices and hence as a constant at least in the short run. Later writers in this tradition have argued that $k_1$ too should be regarded as a function of interest rates (Baumol, 1952; Tobin, 1956).

Although expectations are given great prominence in developing the liquidity function expressing the demand for $M_2$, Keynes and his followers generally did not explicitly introduce an expected interest rate into that function as is done in equation (13). For the most part, in practice, they treated the amount of $M_2$ demanded as a function simply of the current interest rate, the emphasis on expectations serving only as a reason for attributing instability to the liquidity function. Moreover, for the most part, they omitted $P$ (and replaced $y'$ by $Y$) because of their assumption that prices were rigid.

Except for somewhat different language, the analysis up to this point differs from that of earlier quantity theorists, such as Fisher, only by its subtle analysis of the role of expectations about future interest rates, its greater emphasis on current interest rates, and its narrower restriction of the variables explicitly considered as affecting the amount of money demanded.

Keynes's special twist concerned the empirical form of the liquidity-preference function at the low interest rates that he believed would prevail under conditions of underemployment equilibrium. Let the interest rate fall sufficiently low, he argued, and money and bonds would become perfect substitutes for one another; liquidity preference, as he put it, would become absolute. The liquidity-preference function, expressing the quantity of $M_2$ demanded as a function of the rate of interest, would become horizontal at some low but finite rate of interest. Under such circumstances, an increase in the quantity of money by whatever means would lead holders of money to seek to convert their additional cash balances into bonds, which would tend to lower the rate of interest on bonds. Even the slightest lowering would lead speculators with firm expectations to absorb the additional money balances by selling any bonds demanded by the initial holders of the additional money. The result would simply be that the community as a whole would hold the increased quantity of money without any change in the interest rate; $k$ would be higher and $V$ lower. Conversely, a decrease in the quantity of money would lead holders of bonds to seek to restore their money balances by selling bonds, but this would tend to raise the rate of interest, and even the slightest rise would induce the speculators to absorb the bonds offered.

Or, again, suppose nominal income increases or decreases for whatever reason. That will require an increase or decrease in $M_1$ which can come out of or be transferred to $M_2$ without any further effects. The conclusion is that, *under circumstances of absolute liquidity preference,* income can change without a change in $M$ and $M$ can change without a change in income. The holders of money are in metastable equilibrium, like a tumbler on its side on a flat surface; they will be satisfied with whatever the quantity of money happens to be.

Keynes regarded absolute liquidity preference as a strictly "limiting case" of which, though it "might become practically important in future," he knew "of no example … hitherto" (1936, p. 207). However, he treated velocity as if in practice its behaviour frequently approximated that which would prevail in this limiting case.

Keynes's disciples went much farther than Keynes himself. They were readier than he was to accept absolute liquidity preference as the actual state of affairs. More important, many argued that when liquidity preference was not absolute, changes in the quantity of money would affect only the interest rate on bonds and that changes in this interest rate in turn would have little further effect. They argued that both consumption expenditures and investment expenditures were nearly completely insensitive to changes in interest rates, so that a change in $M$ would merely be offset by an opposite and compensatory change in $V$ (or a change in the same direction in $k$), leaving $P$ and $y$ almost completely unaffected. In essence their argument consists in asserting that only paper securities are substitutes for money balances – that real assets never are (see Hansen, 1957, p. 50; Tobin, 1961).

The apparent success during the 1950s and 1960s of governments committed to a Keynesian full-employment policy in achieving rapid economic growth, a high degree of economic stability, and relatively stable prices and interest rates, for a time strongly reinforced belief in the initial Keynesian views about the unimportance of variations in the nominal quantity of money.

The 1970s administered a decisive blow to these views and fostered a revival of belief in the quantity theory. Rapid monetary growth was accompanied not only by accelerated inflation but also by rising, not falling, average levels of unemployment (Friedman, 1977), and by rising, not declining, interest rates. As Robert Lucas put it in 1981,

> Keynesian orthodoxy … appears to be giving seriously wrong answers to the most basic questions of macroeconomic policy. Proponents of a class of models which promised 3½ to 4½ percent unemployment to a society willing to tolerate annual inflation rates of 4 to 5 percent have some explaining to do after a decade [i.e., the 1970s] such as we have just come through. A forecast error of this magnitude and central importance to policy has consequences (pp. 559–60).

This experience undermined the belief that the price level could be regarded as rigid – or at any rate as determined by forces unrelated to the quantity of money; that the nominal quantity of money demanded could be regarded as a function primarily of the nominal interest rate, and that absolute liquidity preference was the normal state of affairs. No teacher of elementary economics since the late 1970s can, as so many did in the 1940s, 1950s, and 1960s, draw on the blackboard a downward sloping liquidity-preference diagram with the nominal quantity of money on the horizontal axis and a nominal interest rate on the vertical axis and confidently proclaim that the only important effect of an increase in the nominal quantity of money would be to lower the rate of interest. The distinction between the nominal interest rate and the real interest rate introduced by Irving Fisher in 1896 has entered – or re-entered – received wisdom (Fisher, 1896).

Despite its subsidence, the Keynesian attack on the quantity theory has left its mark. It has reinforced the tendency, already present in the Cambridge approach, to stress the role of money as an asset and hence to regard the analysis of the demand for money as part of capital or wealth theory, concerned with the composition of the balance sheet or portfolio of assets. The Keynesian stress on autonomous spending and hence on fiscal policy remains important in its own right but also has led to greater emphasis on the effect of government fiscal policies on the demand for money. Keynes's stress on expectations has contributed to the rapid growth in the analysis of the role and formation of expectations in a variety of economic contexts. Conversely, the revival of the quantity theory has led Keynesian economists to treat changes in the quantity of money as an essential element in the analysis of short-term change.

Finally, the controversy between Keynesians and quantity theorists has led both groups to distinguish more sharply between long-run and short-run effects of monetary changes; between "static" or "long-run equilibrium" theory and the dynamics of economic change.

As Franco Modigliani put it in his 1976 presidential address to the American Economic Association, there are currently "no serious analytical disagreements between leading monetarists [i.e., quantity theorists] and leading non-monetarists [i.e., Keynesians]" (1977, p. 1).

However, there still remain important differences on an empirical level. These all centre on the dynamics of short-run change – the process whereby a change in the quantity of money affects aggregate spending and the role of fiscal variables in the process.

The Keynesians regard a change in the quantity of money as affecting in the first instance "the" interest rate, interpreted as a market rate on a fairly narrow class of financial liabilities. They regard spending as affected only "indirectly" as the changed interest rate alters the profitability and amount of investment spending, again interpreted fairly narrowly, and as investment spending, through the multiplier, affects total spending. Hence the emphasis they give in their analysis to the interest elasticities of the demand for money and of investment spending.

The quantity theorists, on the other hand, stress a much broader and more "direct" impact of spending, saying, as in section la above, that individuals will seek "to dispose of what they regard as their excess money balances by paying out a larger sum for the purchase of securities, goods, and services, for the repayment of debts, and as gifts than they are receiving from the corresponding sources."

The two approaches can be readily reconciled on a formal level. Quantity theorists can describe the transmission mechanism as operating "through" the balance sheet and "through" changes in interest rates. The attempt by holders of money to restore or attain a desired balance sheet after an unexpected increase in the quantity of money tends initially to raise the prices of assets and reduce interest rates, which encourages spending to produce new assets and also spending on current services rather than on purchasing existing assets. This is how an initial effect on balance sheets gets translated into an effect on income and spending. The resulting increase in spending tends to raise prices of goods and services which, in turn, by lowering the real value of the quantity of money and of nominal assets, tends to eliminate the initial decline in interest rates, even overshooting in the process.

The difference between the quantity theorists and the Keynesians is less in the nature of the process than in the range of assets considered. The Keynesians tend to concentrate on a narrow range of marketable assets and recorded interest rates. The quantity theorists insist that a far wider range of assets and interest rates must be taken into account – such assets as durable and semi-durable consumer goods, structures, and other real property. As a result, the quantity theorists regard the market rates stressed by the Keynesians as only a small part of the total spectrum of rates that are relevant.

This difference in the assumed transmission mechanism is largely a by-product of the different assumptions about price. The rejection of absolute liquidity preference forced Keynes’s followers to let the interest rate be flexible. This chink in the key assumption that prices are an institutional datum was minimized by interpreting the "interest rate" narrowly, and market institutions made it easy to do so. After all, it is most unusual to quote the "interest rate" implicit in the sales and rental prices of houses and automobiles, let alone furniture, household appliances, clothes, and so on. Hence the prices of these items continued to be regarded as an institutional datum, which forced the transmission process to go through an extremely narrow channel. On the side of the quantity theorists there was no such inhibition. Since they regard prices as flexible, though not "perfectly" flexible, it was natural for them to interpret the transmission mechanism in terms of relative price adjustments over a broad area rather than in terms of narrowly defined interest rates.

Less important differences are the tendency for Keynesians to stress the short-run as opposed to the long-run impact of changes to a far greater extent than the quantity theorists; and, a related difference, to give greater scope to the first-round effect of changes in the quantity of money.

## 3 The Phillips Curve and the Natural Rate Hypothesis

A major postwar development that contributed greatly to the revival of the quantity theory grew out of criticism by quantity theorists of the "Phillips curve" – an allegedly stable inverse relation between unemployment and the rate of change of nominal wages such that a high level of unemployment was accompanied by declining wages, a low level by rising wages. Though not formally linked to the Keynesian theoretical system, the Phillips curve was widely welcomed by Keynesians as helping to fill a gap in the system created by the assumption of rigid wages. In addition, it appeared to offer an attractive trade-off possibility for economic policy: a permanent reduction in the level of unemployment at the cost of a moderate sustained increase in the rate of inflation. The Keynesian assumption that prices and wages could be regarded as institutionally determined made it easy for them to accept a relation between a nominal magnitude (the rate of change of wages) and a real magnitude (unemployment).

By contrast, the quantity theory distinction between real and nominal magnitudes implies that the Phillips curve is theoretically flawed. The quantity of labour demanded is a function of real not nominal wages; and so is the quantity supplied. Under any given set of circumstances, there is an equilibrium level of unemployment corresponding to an equilibrium structure of *real* wage rates. A higher level of unemployment will put downward pressure on real wage rates; a lower level will put upward pressure on real wage rates. The level of unemployment consistent with the equilibrium structure of real wage rates has been termed the "natural rate of unemployment" and defined as

> the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labour and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labour availabilities, the costs of mobility, and so on (Friedman, 1968, p. 8).

The nominal wage rate that corresponds to any given real wage rate depends on the level of prices. Whether that nominal wage rate is rising or falling depends on whether prices are rising or falling. If wages and prices change at the same rate, the real wage rate remains the same. Hence, in the long run, there need be no relation between the rate of change of *nominal* wages and the rate of change of *real* wages, and hence between the rate of change of nominal wages and the level of unemployment. In the long run, therefore, the Phillips curve will tend to be vertical at the natural rate of unemployment – a proposition that came to be termed the Natural Rate Hypothesis.

Over short periods, an *unanticipated* increase in inflation reduces real wages as viewed by employers, inducing them to offer higher nominal wages, which workers erroneously view as higher real wages. This discrepancy simultaneously encourages employers to offer more employment and workers to accept more employment, thereby reducing unemployment, which produces the inverse relation encapsulated in the Phillips curve. However, if the higher rate of inflation continues, the anticipations of workers and employers will converge and the decline in unemployment will be reversed. A negatively sloping Phillips curve is therefore a short-run phenomenon. Moreover, it will not be stable over time, since what matters is not the nominal rates of change of wages and prices but the difference between the actual and the *anticipated* rates of change. The emergence of stagflation in the 1970s quickly confirmed this analysis, leading to the widespread replacement of the original Phillips curve by an expectations-adjusted Phillips curve (Friedman, 1977).

Acceptance of the natural rate hypothesis has had far-reaching effects not only on received wisdom among economists but also on economic policy. It became widely recognized that expansionary monetary and fiscal policies at best gave only a temporary stimulus to output and employment and if long continued would be reflected primarily in inflation.

## 4 The Theory of Rational Expectations

A subsequent theoretical development was the belated flowering of a seed planted in 1961 by John F. Muth, in a long-neglected article on "Rational expectations and the theory of price movements" (Muth, 1961). The theory of rational expectations offers no special insight into stationary-state or long-run equilibrium analysis. Its contribution is to dynamics – short-run change, and hence potentially to stabilization policy.

It has long been recognized by writers of all persuasions that, as Abraham Lincoln put it over a century ago, "you can’t fool all of the people all of the time." The tendency for the public to learn from experience and to adjust to it underlies David Hume’s view that monetary expansion "is favourable to industry" only in its initial stages, but that if it continues, it will come to be anticipated and will affect prices and nominal interest rates but not real magnitudes. It also underlies the companion view associated with the natural rate hypothesis that a "full employment" policy in which monetary, or for that matter fiscal, measures are used to counteract any increase in unemployment will almost inevitably lead not simply to uneven inflation but to uneven inflation around a rising trend – a conclusion often illustrated by analogizing inflation to a drug of which the addict must take larger and larger doses to get the same kick.

Nonetheless, the importance of anticipations and how they are formed in determining the dynamic response to changes in money and other magnitudes remained largely implicit until Lucas and Sargent applied the Muth rational expectations idea explicitly to the reliability of econometric models of the economy and to stabilization policies (Fischer, 1980; Lucas, 1976; Lucas and Sargent, 1981).

The theory of rational expectations asserts that economic agents should be treated as if their anticipations fully incorporate both currently available information about the state of the world and a correct theory of the interrelationships among the variables. Anticipations formed in this way will on the average tend to be correct (a statement whose simplicity conceals fundamental problems of interpretation, Friedman and Schwartz, 1982, pp. 556–7).

The rational expectations hypothesis has far-reaching implications for the validity of econometric models. Suppose a statistician were able to construct a model that predicted highly accurately for a past period all relevant variables; also, that a monetary rule could be devised that if used during the past period with that model could have achieved a particular objective – say keeping unemployment between 4 and 5 per cent. Suppose now that that policy rule were adopted for the future. It would be nearly certain that the model for which the rule was developed would no longer work. The economic equivalent of the Heisenberg indeterminacy principle would take over. The model was for an economy without that monetary rule. Put the rule into effect and it will alter rational expectations and hence behaviour. Even without putting the rule into effect, the model would very likely continue to work only so long as its existence could be kept secret because if market participants learned about it they would use it in forming their rational expectations and thereby falsify it to a greater or lesser extent. Little wonder that every major econometric model is always being sent back to the drawing board as experience confounds it, or that their producers have reacted so strongly to the theory of rational expectations.

The implication of one variant of the theory that has received the most attention and generated the most controversy is the so-called neutrality hypothesis about stabilization policy – in particular, about discretionary monetary policy directed at promoting economic stability Correct rational expectations of economic agents will include correct anticipation of any systematic monetary policy; hence such policy will be allowed for by economic agents in determining their behaviour. Given further the natural rate hypothesis, it follows that any systematic monetary policy will affect the behaviour only of nominal magnitudes and not of such real magnitudes as output and employment. The authorities can affect the course of events only by "fooling" the participants, that is, by acting in an unpredictable, ad hoc way. But, in general, such strictly ad hoc intervention will destabilize the economy, not stabilize it, serving simply to introduce another series of random shocks into the economy to which participants must adapt and which reduce their ability to form precise and accurate expectations.

This is a highly oversimplified account of the rational expectations hypothesis and its implications. All otherwise valid models of the economy will not be falsified by being known. All real effects of systematic and announced governmental policies will not be rendered nugatory. Serious problems have arisen in formulating the hypothesis in a logically satisfactory way, and in giving it empirical content especially in incorporating multi-valued rather than single-valued expectations and allowing for non-independence of events over time. Research in this area is exploding; rapid progress and many changes in received opinion can confidently be anticipated before the rational expectations revolution is fully domesticated.

## 5 Empirical Evidence

There is perhaps no empirical regularity among economic phenomena that is based on so much evidence for so wide a range of circumstances as the connection between substantial changes in the quantity of money and in the level of prices. There are few if any instances in which a substantial change in the quantity of money per unit of output has occurred without a substantial change in the level of prices in the same direction. Conversely, there are few if any instances in which a substantial change in the level of prices has occurred without a substantial change in the quantity of money per unit of output in the same direction. And instances in which prices and the quantity of money have moved together are recorded for many centuries of history, for countries in every part of the globe, and for a wide diversity of monetary arrangements.

The statistical connection itself, however, tells nothing about direction of influence, and this is the question about which there has been the most controversy. A rise or fall in prices, occurring for whatever reason, could produce a corresponding rise or fall in the quantity of money, so that the monetary changes are a passive consequence. Alternatively, changes in the quantity of money could produce changes in prices in the same direction, so that control of the quantity of money implies control of prices. The second interpretation – that substantial changes in the quantity of money are both a necessary and a sufficient condition for substantial changes in the general level of prices – is strongly supported by the variety of monetary arrangements for which a connection between monetary and price movements has been observed. But of course this interpretation does not exclude a reflex influence of changes in prices on the quantity of money. The reflex influence is often important, almost always complex, and, depending on the monetary arrangements, may be in either direction.

*Evidence from specie standards*. Until modern times, money was mostly metallic – copper, brass, silver, gold. The most notable changes in its nominal quantity were produced by sweating and clipping, by governmental edicts changing the nominal values attached to specified physical quantities of the metal, or by discoveries of new sources of specie. Economic history is replete with examples of the first two and their coincidence with corresponding changes in nominal prices (Cipolla, 1956; Feavearyear, 1931). The specie discoveries in the New World in the 16th century are the most important example of the third. The association between the resulting increase in the quantity of money and the price revolution of the 16th and 17th centuries has been well documented (Hamilton, 1934).

Despite the much greater development of deposit money and paper money, the gold discoveries in Australia and the United States in the 1840s were followed by substantial price rises in the 1850s (Cairnes, 1873; Jevons, 1863). When growth of the gold stock slowed, and especially when country after country shifted from silver to gold (Germany in 1871–3, the Latin Monetary Union in 1873, the Netherlands in 1875–6) or returned to gold (the United States in 1879), world prices in terms of gold fell slowly but fairly steadily for about three decades. New gold discoveries in the 1880s and 1890s, powerfully reinforced by improved methods of mining and refining, particularly commercially feasible methods of using the cyanide process to extract gold from low-grade ore, led to much more rapid growth of the world gold stock. Further, no additional important countries shifted to gold. As a result, world prices in terms of gold rose by 25 to 50 per cent from the mid-1890s to 1914 (Bordo and Schwartz, 1984).

*Evidence from great inflations*. Periods of great monetary disturbances provide the most dramatic evidence on the role of the quantity of money. The most striking such periods are the hyperinflations after World War I in Germany, Austria, and Russia, and after World War II in Hungary and Greece, and the rapid price rises, if not hyperinflations, in many South American and some other countries both before and after World War II. These 20th-century episodes have been studied more systematically than earlier ones. The studies demonstrate almost conclusively the critical role of changes in the quantity of money (Cagan, 1965; Meiselman, 1970; Sargent, 1982).

Substantial inflations following a period of relatively stable prices have often had their start in wartime, though recently they have become common under other circumstances. What is important is that something, generally the financing of extraordinary governmental expenditures, produces a more rapid growth of the quantity of money. Prices start to rise, but at a slower pace than the quantity of money, so that for a time the real quantity of money increases. The reason is twofold: first, it takes time for people to readjust their money balances; second, initially there is a general expectation that the rise in prices is temporary and will be followed by a decline. Such expectations make money a desirable form in which to hold assets, and therefore lead to an increase in desired money balances in real terms.

As prices continue to rise, expectations are revised. Holders of money come to expect prices to continue to rise, and reduce desired balances. They also take more active measures to eliminate the discrepancy between actual and desired balances. The result is that prices start to rise faster than the stock of money, and real balances start to decline (that is, velocity starts to rise). How far this process continues depends on the rate of rise in the quantity of money. If it remains fairly stable, real balances settle down at a level that is lower than the initial level but roughly constant – a constant expected rate of inflation implies a roughly constant level of desired real balances; in this case, prices ultimately rise at the same rate as the quantity of money. If the rate of money growth declines, inflation will follow suit, which will in turn lead to an increase in actual and desired real balances as people readjust their expectations; and conversely. Once the process is in full swing, changes in real balances follow with a lag changes in the rate of change of the stock of money. The lag reflects the fact that people apparently base their expectations of future rates of price change partly on an average of experience over the preceding several years, the period of averaging being shorter the more rapid the inflation.

In the extreme cases, those that have degenerated into hyperinflation and a complete breakdown of the medium of exchange, rates of price change have been so high and real balances have been driven down so low as to lead to the widespread introduction of substitute moneys, usually foreign currencies. At that point completely new monetary systems have had to be introduced.

A similar phenomenon has occurred when inflation has been effectively suppressed by price controls, so that there is a substantial gap between the prices that would prevail in the absence of controls and the legally permitted prices. This gap prevents money from functioning as an effective medium of exchange and also leads to the introduction of substitute moneys, sometimes rather bizarre ones like the cigarettes and cognac used in post-World War II Germany.

*Other evidence*. The past two decades have witnessed a literal flood of literature dealing with monetary phenomena. Expressed in broad terms, the literature has been of two overlapping types – qualitative and econometric – and has dealt with two overlapping sets of issues – static or long-term effects of monetary change and dynamic or cyclical effects.

Some broad findings are:

(1) For both long and short periods there is a consistent though not precise relation between the rate of growth of the quantity of money and the rate of growth of nominal income. If the quantity of money grows rapidly, so will nominal income, and conversely. This relation is much closer for long than for short periods.

Two recent econometric studies have tested the long-run effects using comparisons among countries for the post-World War II period. Lothian concludes his study for 20 countries for the period 1956–80:

> In this paper I have examined three sets of hypotheses associated with the quantity theory of money: the classical neutrality proposition [i.e., changes in the nominal quantity of money do not affect real magnitudes in the long run], the monetary approach to exchange rates [i.e., changes in exchange rates between countries reflect primarily changes in money per unit of output in the several countries], and the Fisher equation [i.e., differences in sustained rates of inflation produce corresponding differences in nominal interest rates]. The data are completely consistent with the first two and moderately supportive of the last (1985, p. 835).

Duck concludes his study for 33 countries and the period 1962 to 1982 – which uses overlapping data but substantially different methods:

> Its [the study's] findings suggest that (i) the real demand for money is reasonably well explained by a small number of variables, principally real income and interest rates; (ii) nominal income is closely related to the quantity of money, but is also related to the behaviour of other variables, principally interest rates; (iii) most changes in nominal income or its determinants are absorbed by price increases; (iv) even over a 20-year period some nominal income growth is to a significant degree absorbed by real output growth; (v) the evidence that expectations are rational is weak (1985, p. 33).

(2) These findings for the long run reflect a long-run real demand function for money involving, as Duck notes, a small number of variables, that is highly stable and very similar for different countries. The elasticity of this function with respect to real income is close to unity, occasionally lower, generally higher, especially for countries that are growing rapidly and in which the scope of the money economy is expanding. The elasticity with respect to interest rates is, as expected, negative but relatively low in absolute value. The real quantity demanded is not affected by the price level (i.e., there is no "monetary illusion") (Friedman and Schwartz, 1982; Laidler, 1985).

(3) Over short periods, the relation between growth in money and in nominal income is often concealed from the naked eye partly because the relation is less close for short than long periods but mostly because it takes time for changes in monetary growth to affect income, and how long it takes is itself variable. Today's income growth is not closely related to today's monetary growth; it depends on what has been happening to money in the past. What happens to money today affects what is going to happen to income in the future.

(4) For most major Western countries, a change in the rate of monetary growth produces a change in the rate of growth of nominal income about six to nine months later. This is an average that does not hold in every individual case. Sometimes the delay is longer, sometimes shorter. In particular, it tends to be shorter under conditions of high and highly variable rates of monetary growth and of inflation.

(5) In cyclical episodes the response of nominal income, allowing for the time delay, is greater in amplitude than the change in monetary growth, so that velocity tends to rise during the expansion phase of a business cycle and to fall during the contraction phase. This reaction appears to be partly a response to the pro-cyclical pattern of interest rates; partly to the linkage of desired cash balances to permanent rather than measured income.

(6) The changed rate of growth of nominal income typically shows up first in output and hardly at all in prices. If the rate of monetary growth increases or decreases, the rate of growth of nominal income and also of physical output tends to increase or decrease about six to nine months later, but the rate of price rise is affected very little.

(7) The effect on prices, like that on income and output, is distributed over time, but comes some 12 to 18 months later, so that the total delay between a change in monetary growth and a change in the rate of inflation averages something like two years. That is why it is a long row to hoe to stop an inflation that has been allowed to start. It cannot be stopped overnight.

(8) Even after allowance for the delayed effect of monetary growth, the relation is far from perfect. There’s many a slip over short periods ‘twixt the monetary change and the income change.

(9) In the short run, which may be as long as three to ten years, monetary changes affect primarily output. Over decades, on the other hand, as already noted, the rate of monetary growth affects primarily prices. What happens to output depends on real factors: the enterprise, ingenuity and industry of the people; the extent of thrift; the structure of industry and government; the relations among nations, and so on. (In re points 3 to 9, Friedman and Schwartz, 1963a, 1963b; Friedman, 1961, 1977, 1984; Judd and Scadding, 1982.)

(10) One major finding has to do with severe depressions. There is strong evidence that a monetary crisis, involving a substantial decline in the quantity of money, is a necessary and sufficient condition for a major depression. Fluctuations in monetary growth are also systematically related to minor ups and downs in the economy, but do not play as dominant a role compared to other forces. As Friedman and Schwartz put it,

> Changes in the money stock are … a consequence as well as an independent source of change in money income and prices, though, once they occur, they produce in their turn still further effects on income and prices. Mutual interaction, but with money rather clearly the senior partner in longer-run movements and in major cyclical movements, and more nearly an equal partner with money income and prices in shorter-run and milder movements – this is the generalization suggested by our evidence (1963b, p. 695; Friedman and Schwartz, 1963a; Cagan, 1965, pp. 296–8).

(11) A major unsettled issue is the short-run division of a change in nominal income between output and price. The division has varied widely over space and time and there exists no satisfactory theory that isolates the factors responsible for the variability (Gordon, 1980, 1981, 1982; Friedman and Schwartz, 1982, pp. 59–62).

(12) It follows from these propositions that *inflation is always and everywhere a monetary phenomenon* in the sense that it is and can be produced only by a more rapid increase in the quantity of money than in output. Many phenomena can produce temporary fluctuations in the rate of inflation, but they can have lasting effects only insofar as they affect the rate of monetary growth. However, there are many different possible reasons for monetary growth, including gold discoveries, financing of government spending, and financing of private spending. Hence, these propositions are only the beginning of an answer to the causes and cures for inflation. The deeper question is why excessive monetary growth occurs.

(13) Government spending may or may not be inflationary. It clearly will be inflationary if it is financed by creating money, that is, by printing currency or creating bank deposits. If it is financed by taxes or by borrowing from the public, the main effect is that the government spends the funds instead of the taxpayer or instead of the lender or instead of the person who would otherwise have borrowed the funds. Fiscal policy is extremely important in determining what fraction of total national income is spent by government and who bears the burden of that expenditure. It is also extremely important in determining monetary policy and, via that route, inflation. Essentially all major inflations, especially hyperinflations, have resulted from resort by governments to the printing press to finance their expenditures under conditions of great stress such as defeat in war or internal revolution, circumstances that have limited the ability of governments to acquire resources through explicit taxation.

(14) A change in monetary growth affects interest rates in one direction at first but in the opposite direction later on. More rapid monetary growth at first tends to lower interest rates. But later on, the resulting acceleration in spending and still later in inflation produces a rise in the demand for loans which tends to raise interest rates. In addition, higher inflation widens the difference between real and nominal interest rates. As both lenders and borrowers come to anticipate inflation, lenders demand, and borrowers are willing to offer, higher nominal rates to offset the anticipated inflation. That is why interest rates are highest in countries that have had the most rapid growth in the quantity of money and also in prices – countries like Brazil, Chile, Israel, South Korea. In the opposite direction, a slower rate of monetary growth at first raises interest rates but later on, as it decelerates spending and inflation, lowers interest rates. That is why interest rates are lowest in countries that *have had* the slowest rate of growth in the quantity of money – countries like Switzerland, Germany, and Japan.

(15) In the major Western countries, the link to gold and the resultant long-term predictability of the price level meant that until some time after World War II, interest rates behaved as if prices were expected to be stable and both inflation and deflation were unanticipated; the so-called Fisher effect was almost completely absent. Nominal returns on nominal assets were relatively stable; real returns unstable, absorbing almost fully inflation and deflation.

(16) Beginning in the 1960s, and especially after the end of Bretton Woods in 1971, interest rates started to parallel rates of inflation. Nominal returns on nominal assets became more variable; real returns on nominal assets, less variable (Friedman and Schwartz, 1982, pp. 10–11).

## 6. Policy Implications

On a very general level the implications of the quantity theory for economic policy are straightforward and clear. On a more precise and detailed level they are not.

Acceptance of the quantity theory means that the quantity of money is a key variable in policies directed at controlling the level of prices or of nominal income. Inflation can be prevented if and only if the quantity of money per unit of output can be kept from increasing appreciably. Deflation can be prevented if and only if the quantity of money per unit of output can be kept from decreasing appreciably. This implication is by no means trivial. Monetary authorities have Congress to establish a Commission on the Role of Gold. In its final report, "the Commission concludes that, under present circumstances, restoring a gold standard does not appear to be a fruitful method for dealing with the continuing problem of inflation…. We favour no change in the flexible exchange rate system" (Commission, 1982, vol. 1, pp. 17, 20). The testimony before the Commission revealed that agreement on a "gold standard" concealed wide differences in the precise meaning of the phrase, varying from a system in which money consisted of full-bodied gold or warehouse receipts for gold to one in which the monetary authorities were instructed to regard the price of gold as one factor affecting their policy.

A very different component of the discussion has to do with possible alternatives to gold as a long-term anchor to the price level. This includes proposals for subjecting monetary authorities to more specific legislative or constitutional guidelines, varying from guidelines dealing with their objectives (price stability, rate of growth of nominal income, real interest rate, etc.) to guidelines specifying a specific rate of growth in money or high-powered money. Perhaps the most widely discussed proposal along this line is the proposal for imposing on the authorities the obligation to achieve a constant rate of growth in a specified monetary aggregate (Friedman, 1960, pp. 92–5; Commission, 1982, vol. 1, p. 17). Other proposals include freezing the stock of base money and eliminating discretionary monetary policy, and denationalizing money entirely, leaving it to the private market and a free banking system (Friedman, 1984; Friedman and Schwartz, 1986; Hayek, 1976; White, 1984a).

Finally, a still more radical series of proposals is that the unit of account be separated from the medium of exchange function, in the belief that financial innovation will establish an efficient payment system dispensing entirely with the use of cash. The specific proposals are highly sophisticated and complex, and have been sharply criticized. So far, their value has been primarily as a stimulus to a deeper analysis of the meaning and role of money. (For the proposals, see Black, 1970; Fama, 1980; Hall, 1982a, 1982b; Greenfield and Yeager, 1983; for the criticisms, see White, 1984b; McCallum, 1985).

One thing is certain: the quantity theory of money will continue to generate agreement, controversy, repudiation, and scientific analysis, and will continue to play a role in government policy during the next century as it has for the past three.

## Bibliography

Angell, J.W. 1936. *The Behavior of Money*. New York: McGraw-Hill.

Bagehot, W. 1873. *Lombard Street*. London: Henry S. King.

Barnett, W.A., Offenbacher, E.K. and Spindt, P.A. 1984. The new Divisia monetary aggregates. *Journal of Political Economy* 92(6), December, 1049–85.

Baumol, W.J. 1952. The transactions demand for cash: an inventory theoretic approach. *Quarterly Journal of Economics* 66, November, 545–56.

Black, F. 1970. Banking and interest rates in a world without money: the effects of uncontrolled banking, *Journal of Bank Research* 1(3), Autumn, 2–20.

Black, H. 1975. The relative importance of determinants of the money supply: the British case. *Journal of Monetary Economics* 1(2), April, 25–64.

Bordo, M.D. and Schwartz, A.J. (eds) 1984. *A Retrospective on the Classical Gold Standard, 1821–1931*. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Burger, A.E. 1971. *The Money Supply Process*. Belmont: Wadsworth.

Cagan, P. 1965. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia University Press for the National Bureau of Economic Research.

Cagan, P. 1972. *The Channels of Monetary Effects on Interest Rates*. New York: National Bureau of Economic Research.

Cairnes, J.E. 1873. Essays on the gold question. In J.E. Cairnes, *Essays in Political Economy*, London: Macmillan.

Cipolla, C.M. 1956. *Money, Prices, and Civilization in the Mediterranean World, Fifth to Seventeenth Century*. Princeton: Princeton University Press.

Commission on the Role of Gold in the Domestic and International Monetary Systems. 1982. *Report to the Congress*, March. Washington, D.C.: The Commission.

Duck, N.W. 1985. Money, output and prices: an empirical study using long-term cross country data. Working Paper, University of Bristol, September.

Fama, E.F. 1980. Banking in the theory of finance. *Journal of Monetary Economics* 6(1), January, 39–57.

Feavearyear, A.E. 1931. *The Pound Sterling: a History of English Money*. 2nd edn, Oxford: Clarendon Press, 1963.

Fischer, S. (ed.) 1980. *Rational Expectations and Economic Policy*. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Fisher, I. 1896. *Appreciation and Interest*. Ner York: American Economic Association.

Fisher, I. 1911. *The Purchasing Power of Money*. 2nd revised edn, 1926; reprinted New York: Kelley, 1963.

Fisher, I. 1919. Money, prices, credit and banking. *American Economic Review* 9, June, 407–9.

Frenkel, J.A. 1976. Adjustment mechanisms and the monetary approach to the balance of payments. In *Recent Issues in International Monetary Economics*, ed. E. Claassen and P. Salin, Amsterdam: North-Holland.

Frenkel, J.A. and Johnson, H.G. 1976. The monetary approach to the balance of payments: essential concepts and historical origins. In *The Monetary Approach to the Balance of Payments*, ed. J.A. Frenkel and H.G. Johnson, Toronto: University of Toronto Press.

Friedman, M. 1956. The quantity theory of money – a restatement. In *Studies in the Quantity Theory of Money*, ed. M. Friedman, Chicago: University of Chicago Press.

Friedman, M. 1957. *A Theory of the Consumption Function*. Princeton: Princeton University Press for the National Bureau of Economic Research.

Friedman, M. 1959. The demand for money: some theoretical and empirical results. *Journal of Political Economy* 67, August, 327–51. Reprinted as Occasional Paper No. 68, New York: National Bureau of Economic Research, and in Friedman (1969).

Friedman, M. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, M. 1961. The lag in effect of monetary policy. *Journal of Political Economy* 69, October, 447–66. Reprinted in Friedman (1969).

Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), March, 1–17. Reprinted in Friedman (1969).

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine.

Friedman, M. 1977. Inflation and unemployment (Nobel lecture). *Journal of Political Economy* 85(3), June, 451–72.

Friedman, M. 1984. Monetary policy for the 1980s. In *To Promote Prosperity: U.S. domestic policy in the mid-1980s*, ed. J.H. Moore, Stanford: Hoover Institution Press.

Friedman, M. and Schwartz, A.J. 1963a. Money and business cycles. *Review of Economics and Statistics* 45(1), Supplement, February, 32–64. Reprinted in Friedman (1969).

Friedman, M. and Schwartz, A.J. 1963b. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1970. *Monetary Statistics of the United States*. New York: Columbia University Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1982. *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Friedman, M. and Schwartz, A.J. 1986. Has government any role in money? *Journal of Monetary Economics* 17(1), January, 37–62.

Gordon, R.J. 1980. A consistent characterization of a near-century of price behavior. *American Economic Review* 70(2), May, 243–49.

Gordon, R.J. 1981. Output fluctuations and gradual price adjustment. *Journal of Economic Literature* 19(2), June, 493–530.

Gordon, R.J. 1982. Price inertia and policy ineffectiveness in the United States, 1890–1980. *Journal of Political Economy* 90(6), December, 1087–117.

Greenfield, R.L. and Yeager, L.B. 1983. A laissez-faire approach to monetary stability. *Journal of Money, Credit, and Banking* 15(3), August, 302–15.

Haberler, G. 1941. *Prosperity and Depression*. 3rd edn, Geneva: League of Nations.

Hall, R.E. 1982a. Explorations in the gold standard and related policies for stabilizing the dollar. In *Inflation: Causes and Effects*, ed. R.E. Hall, Chicago: University of Chicago Press.

Hall, R.E. 1982b. ‘Monetary trends in the United States and the United Kingdom’: a review from the perspective of new developments in monetary economics. *Journal of Economic Literature* 20(4), December, 1552–6.

Hamilton, E.J. 1934. *American Treasure and the Price Revolution in Spain, 1501–1650*. Harvard Economic Studies, Vol. 43, New York: Octagon, 1965.

Hansen, A. 1957. *The American Economy*. New York: McGraw-Hill.

Hayek, F.A. 1976. *Denationalization of Money*. 2nd extended edn, London: Institute of Economic Affairs, 1978.

Hume, D. 1752. Of interest; of money. In *Essays, Moral, Political and Literary*, Vol. 1 of *Essays and Treatises*, a new edn, Edinburgh: Bell and Bradfute, Cadell and Davies, 1804.

Humphrey, T.M. 1984. Algebraic quantity equations before Fisher and Pigou. *Economic Review*, Federal Reserve Bank of Richmond 70(5), September-October, 13–22.

Ihori, T. 1985. On the welfare cost of permanent inflation. *Journal of Money, Credit, and Banking* 17(2), May, 220–31.

Jevons, W.S. 1863. A serious fall in the value of gold. In *Investigations in Currency and Finance*, 2nd edn, London: Macmillan, 1909.

Johnson, H.G. 1961. *The General Theory* after twenty-five years. *American Economic Association, Papers and Proceedings* 51, May, 1–17.

Judd, J.P. and Scadding, J.L. 1982. The search for a stable money demand function. *Journal of Economic Literature* 20(3), September, 993–1023.

Keynes, J.M. 1923. *A Tract on Monetary Reform*. Reprinted London: Macmillan for the Royal Economic Society, 1971.

Keynes, J.M. 1936. *The General Theory of Employment, Interest, and Money*. Reprinted London: Macmillan for the Royal Economic Society, 1973.

Knoester, A. 1984. Pigou and buffer effects in monetary economics. Discussion Paper 8406 G/M, Institute for Economic Research, Erasmus University, Rotterdam.

Laidler, D. 1984. The ‘buffer stock’ notion in monetary economics. *Economic Journal* 94, Supplement, 17–34.

Laidler, D. 1985. *The Demand for Money: theories, evidence, and problems*. 3rd edn, New York: Harper & Row.

Lothian, J.R. 1985. Equilibrium relationships between money and other economic variables. *American Economic Review* 75(4), September, 828–35.

Lucas, R.E., Jr. 1976. Econometric policy evaluation: a critique. *Journal of Monetary Economics* supplementary series 1, 19–46.

Lucas, R.E., Jr. 1981. Tobin and monetarism: a review article. *Journal of Economic Literature* 19(2), June, 558–67.

Lucas, R.E., Jr. and Sargent, T.J. (eds.) 1981. *Rational Expectations and Economic Practice*. 2 vols, Minneapolis: University of Minnesota Press.

McCallum, B. 1985. Bank deregulation, accounting systems of exchange and the unit of account: a critical review. *Carnegie-Rochester Conference Series on Public Policy* 23, Autumn.

McKinnon, R. 1984. *An International Standard for Monetary Stabilization*. Cambridge, Mass: MIT Press.

Meiselman, D. (ed.) 1970. *Varieties of Monetary Experience*. Chicago: University of Chicago Press.

Mill, J.S. 1844. Review of books by Thomas Tooke and R. Torrens. *Westminster Review*, June.

Miller, M.H. and Orr, D. 1966. A model of the demand for money by firms. *Quarterly Journal of Economics* 80(3), August, 413–35.

Miller, M.H. and Orr, D. 1968. The demand for money by firms: extensions of analytical results. *Journal of Finance* 23(5), December, 735–59.

Modigliani, F. 1977. The monetarist controversy, or should we forsake stabilization policies? *American Economic Review* 67(2), March, 1–19.

Mussa, M. 1977. The welfare cost of inflation and the role of money as a unit of account. *Journal of Money, Credit, and Banking* 9(2), May, 276–86.

Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29, July, 315–35. Reprinted in Lucas and Sargent (1981).

Newcomb, S. 1885. *Principles of Political Economy*. New York: Harper & Brothers.

Patinkin, D. 1948. Price flexibility and full employment. *American Economic Review* 38, September, 543–64. Revised and reprinted in F.A. Lutz and L.W. Mints (American Economic Association), *Readings in Monetary Theory*, Homewood, Ill.: Irwin, 1951.

Phelps, E.S. 1967. Phillips curves, expectations of inflation, and optimal unemployment over time. *Economica* 34(135), August, 254–81.

Pigou, A.C. 1917. The value of money. *Quarterly Journal of Economics* 32, November, 38–65. Reprinted in F.A. Lutz and L.W. Mints (American Economic Association), *Readings in Monetary Theory*, Homewood, Ill.: Irwin, 1951.

Pigou, A.C. 1947. Economic progress in a stable environment. *Economica* 14(55), August, 180–88.

Sargent, T.J. 1982. The ends of four big inflations. In *Inflation: Causes and Effects*, ed. R.E. Hall, Chicago: University of Chicago Press.

Snyder, C. 1934. On the statistical relation of trade, credit, and prices. *Revue de l'Institut International de Statistique* 2, October, 278–91.

Spindt, P.A. 1985. Money is what money does: monetary aggregation and the equation of exchange. *Journal of Political Economy* 93(1), February, 1975–2204.

Tobin, J. 1947. Money wage rates and employment. In *The New Economics*, ed. S. Harris, New York: Knopf.

Tobin, J. 1956. The interest-elasticity of transactions demand for cash. *Review of Economics and Statistics* 38, August, 241–47.

Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25, February, 65–86.

Tobin, J. 1961. Money, capital and other stores of value. *American Economic Review, Papers and Proceedings* 51, May, 26–37.

Tobin, J. 1974. Friedman's theoretical framework. In *Milton Friedman's Monetary Framework: a Debate with His Critics*, ed. R.J. Gordon, Chicago: University of Chicago Press.

White, L.H. 1984a. *Free Banking in Britain: Theory, Experience and Debate, 1800–1845*. New York: Cambridge University Press.

White, L.H. 1984b. Competitive payments systems and the unit of account. *American Economic Review* 74(4), September, 699–712.

---

Adaptation, entitled "Notes on the Quantity Theory of Money," printed in Kurt R. Leube, editor, *The Essence of Friedman*, pp. 370-384. Stanford, California: Hoover Institution, 1987.

This is a post-peer-review, pre-copyedit version of an article published in *The New Palgrave: A Dictionary of Economics*. The definitive publisher-authenticated version [Friedman, Milton. "Quantity Theory of Money." *The New Palgrave: A Dictionary of Economics*. First Edition. Eds. John Eatwell, Murray Milgate and Peter Newman. Palgrave Macmillan, 1987.] is available online at: http://www.dictionaryofeconomics.com/article?id=pde1987_X001813

11/25/13