1974

The Optimal Timing of Transactions: An Extension of the Inventory Theory of the Demand for Money

Peter Howitt

Robert W. Clower

Follow this and additional works at: https://ir.lib.uwo.ca/economicsresrpt

Part of the Economics Commons

Citation of this paper:
RESEARCH REPORT 7406

THE OPTIMAL TIMING OF TRANSACTIONS:
AN EXTENSION OF THE INVENTORY THEORY
OF THE DEMAND FOR MONEY

P. W. Howitt and R. W. Clower
THE OPTIMAL TIMING OF TRANSACTIONS:

AN EXTENSION OF THE INVENTORY THEORY OF THE DEMAND FOR MONEY

by

P. W. Howitt and R. W. Clower*

This paper reports some results of ongoing research that appear to challenge prevailing ideas about the inventory theory of the demand for money. The analysis extends earlier work by treating the timing of both sales and purchases as individual decision variables, thereby relaxing conventional a priori restrictions on the relative frequency of transactions. This procedure permits us to shed new light on the relation between commodity inventories and holdings of money, to clarify the circumstances in which individuals will choose to hold positive transactions balances, and to establish more precisely the nature of potential gains associated with payments of interest on money. In the course of the argument we also derive and discuss some comparative statics results and comment briefly on collateral literature.

I. BACKGROUND AND PROCEDURE

Recent discussions of the logistics and optimality of monetary exchange have directed the attention of economists to several important but previously unappreciated weaknesses in the microfoundations of monetary theory. The standard analytical device of treating real money balances as an argument in utility and production functions serves formally to generate a demand for money, but it simultaneously rules out deeper analysis of the economics of monetary exchange. In particular, it precludes any but elliptical analysis of the "services of money"\(^1\) and so begs rather than answers the question why money should ever be held at all.\(^2\) The obvious alternative to the traditional approach is to link the services of money directly with exchange activities, but this poses other problems since it requires that exchange activities be characterized more precisely than is customary in established theory.\(^3\)
The difficulties inherent in the latter approach may be appreciated by considering recent attempts to use the familiar Baumol-Tobin inventory theory of financial transactions to analyze transactions involving real commodities. As in the Baumol-Tobin theory, so in this literature the individual trader is assumed to decide the frequency of transactions by balancing holding costs against costs of transacting; the only novelty lies in the substitution of commodity transactions for bond transactions. The trader's holdings of commodity inventories are automatically fixed by his timing of purchases, and average money balances are then determined by the requirement that holdings at some initial time be just large enough to finance all subsequent purchases. The holding of positive money balances thus emerges as a mechanical consequence of the decision not to synchronize purchases perfectly with sales. Although this procedure represents a considerable extension of the transactions approach to the demand for money, it permits no more than a partial analysis of the problem because it takes the timing of sales as given. This would be of minor importance if we were concerned exclusively with household transactors (for which it may be a good first approximation to suppose that the timing of receipts is given). For some households and for most other kinds of transactors, however, the timing of sales is as much a decision variable as the timing of purchases, so it is desirable to handle the two cases symmetrically (cf. Barro [1973]). The usual inventory theory approach also imposes arbitrary a priori restrictions on relative frequencies of purchase and sale by supposing not only that the frequency of sales is less than that of purchases, but also that the frequency of purchases is an exact multiple of the frequency of sales, i.e., that the ratio of the frequency of purchases to the frequency of sales is an integer. Even though this ratio is eventually treated as a continuous variable, the possibility of non-integer values is not considered in the formulation of the decision problem.
The present analysis differs from earlier work mainly in being more precise and general in its treatment of fundamentals. Following earlier writers, we start by imagining an economy in which just one commodity, called money, is routinely accepted in exchange for other commodities in all organized markets. We suppose that, in executing any purchase or sale transaction, individual traders incur positive set-up costs (i.e., costs that are to some extent fixed, so that costs per unit of time vary inversely with quantities purchased or sold per transaction). This means that individuals will trade, if at all, at discrete points rather than continuously in time, which implies, in turn, that each active trader will, in general, hold positive stocks of all goods purchased for consumption and all commodities produced and sold. However, since nothing in the logic of the situation prevents any trader from so choosing dates of purchase and sale that money receipts and outlays are perfectly synchronized, we cannot assert that every active trader will on average hold positive stocks of money. To say anything definite about that question, we must first establish precise links between individual holdings of money balances and alternative patterns of commodity purchase and sale.

II. THE MONEY REQUIREMENTS FUNCTION

To avoid obscuring fundamental ideas in a haze of extraneous complications, we proceed initially by considering an individual trader who produces and sells units of just one commodity (S) and purchases and consumes units of just one other (D). In this special case, a complete description of observable behavior may be provided, in principle, by specifying just four time functions:

(i) the continuous time path of production, \( s(t), t \in \mathbb{R} \)

(ii) the continuous time path of consumption, \( d(t), t \in \mathbb{R} \)

(iii) the discrete time sequence of sales (money receipts), \( \{S_{i+1}\}_{i \in I} \)
(iv) the discrete time sequence of purchases (money outlays),

\[
\{D_{v_j}\}_{j \in I},
\]

where \(R\) and \(I\) denote, respectively, the set of all real numbers and the set of all integers, \(t\) is a continuous time variable and \(u_i\) and \(v_j\) denote sale and purchase dates (i.e., \(u_i\) and \(v_j\) range over discrete subsets of \(R\)). We treat all market prices as given and choose units of measurement to make the money prices of all commodities unity so that we need not distinguish notationally or otherwise between real and money magnitudes.

Explicit dynamic modelling of the individual's behavior for this special case, although possible in principle, is not attempted here, partly because the task is far more formidable than might be supposed,\(^7\) partly because the worth of such an exercise--given the assured complexity of its results--is problematical. In what follows, we deal with less ambitious but immediately more revealing conceptual experiments that permit us explicitly to characterize stationary solutions to dynamic models that themselves remain implicit in the background. The fruitfulness of this procedure is clearly documented in the history of economics and other sciences.

For our purposes, no loss of essential generality is involved in supposing that the time path of output is given and constant. Our basic criterion for stationarity may then be expressed by the condition that expenditure at all times be equal to income; that is

\[(S.1) \quad d(t) = s(t) = s \quad \text{for all } t.\]

We shall further restrict the possible transaction sequences by requiring that all purchases be of the same amount and equally spaced, and that the same be true of all sales.
\[
\begin{align*}
S_{u_i} &= S, \quad u_i - u_{i-1} = \Delta u, \quad \text{for all } i; \quad \text{and} \\
D_{v_j} &= D, \quad v_j - v_{j-1} = \Delta v, \quad \text{for all } j.
\end{align*}
\]

It involves no loss of generality to suppose further that purchases occur only when the consumption good has been exhausted, and that whenever a sale occurs the entire accumulated inventory of the production good is sold. This implies that:

\[(S.3) \quad \Delta u = S/s, \quad \Delta v = D/s.\]

So far we have said nothing about the phasing of transactions, just their size and frequency. For example, we might have a trader buying one unit once a week and also selling one unit once a week, but this doesn't tell us whether the two weekly transactions always occur at the same time during the week, or whether, for example, the purchase always occurs twenty-four hours after (or even before) the sale. We might suppose that there is a cost of bunching the transactions too closely together. On the other hand there will clearly be indirect costs of having the transactions spread too far apart because the longer a purchase is postponed the larger will be the trader's average money holdings, and thus the larger will be the interest-opportunity cost. A complete account of the timing of transactions would allow the transactor to determine the optimal phasing as well as frequency and size. However the question of phasing introduces several complications that are most easily dealt with separately from the analysis of size and frequency. Thus we shall proceed initially with the simplifying assumption that there are no bunching costs, in which case the trader will always choose the phasing that minimizes his average money holdings. This may be expressed very simply by the assumption that
(S.4) There exists at least one date, \( t_0 \), at which a purchase and sale coincide.

We refer to any set of functions (i) \( \sim \) (iv) that satisfy (S.1) \( \sim \) (S.4) as a stationary trading process. Any stationary trading process is uniquely characterized by the two variables \( D \) and \( S \). Given any stationary trading process we may characterize the time path of the trader's money holdings as:

\[
M(t) = M(t_0) + \sum_{u_i} S_{u_i} - \sum_{v_j} D_{v_j}, \quad \text{for } t > t_0.
\]

\( t_0 \leq u_i \leq t \quad t_0 < v_j \leq t \)

This time path depends upon the amount of money, \( M(t_0) \), held just after the simultaneous purchase and sale at \( t_0 \). All that remains for us to do now is to determine the value of \( M(t_0) \).

We assume that the trader is not allowed to let his money balances fall below an amount \( -L \) (where \( L \geq 0 \)), which may be interpreted as his overdraft limit. There is no technical reason why this limit must ever be binding; however, as long as the rate of interest charged on overdrafts is less than the interest-opportunity cost of holding positive money balances, it will clearly be in the trader's interest to make sure that his money holdings are minimized subject to the overdraft constraint. In order to avoid unnecessary complications we shall limit our analysis to this case, and make the assumption that the trader chooses the smallest value of \( M(t_0) \) such that \( M(t) \geq -L \) for all \( t \). The implied time path will be referred to as the required path of money.

In order to construct the required path, suppose provisionally that the trader starts just before date \( t_0 \) with zero money balances. That is, \( M(t_0) = S - D \). Then the implied virtual path will be as illustrated by the dashed line in Figure 1. The virtual path will reach zero at some date, and unless \( D \) is an exact divisor of \( S \) the virtual path will become negative. The
Inventory time-paths when $s = 1$, $D = 2$, $S = 3$, $L = \frac{1}{2}$.

Figure 1
minimum value of the virtual path is denoted by \( -V \) (where \( V \geq 0 \)). The required path can be constructed from the virtual path by raising or lowering the whole path until this minimum value equals \( -L \). That is, we must add the amount \( V - L \) to the virtual path, so that it is given by the difference equation above, with \( M(t_o) = S - D + V - L \).

At date \( t_o \) the trader's inventories will consist of this amount of money plus the amount \( D \) of the consumption good. Thus the total value of his inventories will be \( S + V - L \). By virtue of the stationarity hypothesis this must also be the total value of inventories at all points in time. It follows that

\[
\overline{M} + \overline{D} + \overline{S} = S + V - L
\]  

(1)

where \( \overline{D} = D/2 \) and \( \overline{S} = S/2 \) are the average holdings of the two non-money inventories, and \( \overline{M} \) is the average holding of money balances. It is shown in the mathematical appendix that

\[
V = D - G(D, S)
\]  

(2)

where \( G(D, S) \) is equal to the greatest common divisor of \( D \) and \( S \) if \( D/S \) is rational,\(^{10}\) and \( G(D, S) = 0 \) otherwise. This function plays a fundamental role in the whole of the ensuing analysis, and it is not very well-behaved in the usual sense. For later reference, therefore, we list here some of its more important properties:\(^{11}\)

(P.1) \( G(\lambda D, \lambda S) = \lambda G(D, S) \) for any \( \lambda > 0 \); i.e., the function is homogeneous of degree one in both arguments.

(P.2) \( 0 \leq G(D, S) \leq \min(D, S) \).

(P.3) \( G(D, S) = \begin{cases} 
0 \text{ if } D/S \text{ is irrational} \\
\min(D, S) \text{ if } D \text{ is an exact multiple or divisor of } S.
\end{cases} \)
(P.4) \(G(D,S)\) has jump-discontinuities everywhere that \(D/S\) is rational. More precisely, \(G(D,S) > 0\) for such points, but if any sequence \(\{D^q, S^q\}\) approaches such a point, and \(D^q/S^q \neq D/S\) for all \(q\), then \(G(D^q, S^q)\) approaches zero.

The first three properties are easy to verify. The fourth is not so obvious, but a simple proof is included in the mathematical appendix.

Combining (1) and (2),

\[
\bar{M} = \bar{D} + \bar{S} - G(D,S) - L
\]

(3)

The finance requirement of a stationary trading process is defined as the sum of average money holdings plus overdraft limit, \(F = \bar{M} + L\). From (3), this is given by the finance function:

\[
F(D,S) = \frac{1}{g}(D+S) - G(D,S).
\]

(4)

Figure 2 is an attempt to show how \(F\) varies as \(D\) is varied with \(S\) held constant. When \(D=S\), \(F=0\). Whenever \(D\) is an exact multiple or divisor of \(S\), \(F\) equals \(|D-S|\), and lies on the broken line ASz, as indicated by the points a,b,c,... and x,y,z,... This broken line forms the lower boundary of the function.

When \(D/S\) is irrational, \(F\) lies on the upper boundary, given by the line \(AB(F=\bar{D}+\bar{S})\). When \(D/S\) is rational, but \(D\) is neither an exact multiple nor an exact divisor of \(S\), then \(F\) lies strictly between these two boundaries.

It follows from (P.4) that for any given \(S\), \(F\) will attain a discontinuous local minimum at every value of \(D\) for which \(D/S\) is rational. This means that for such values of \(D\), \(F\) consists of an infinity of isolated points. However, the most important feature of the finance function is the discontinuous nature of the local minima defined by values of \(D\) that are exact multiples or divisors of \(S\). These particular discontinuities are of potential significance because their effect clearly will be to discourage the trader from choosing values of \(D\) and \(S\) that do not correspond to points
The finance function

F = D + S

Figure 2
on the lower boundary of the finance function.

The economic interpretation of these discontinuities is straightforward. Suppose, for example, that we start with \(L=0\) and \(D=S\), in which case purchases will be perfectly synchronized with sales and no money holdings will be required. In particular, the trader may start at \(t_0\) with no money-holdings. But if \(D\) is now reduced by a small amount, with \(S\) held constant, the trader must start at \(t_0\) with money holdings equal to \(D\), because his second purchase will occur just prior to his second sale.

III. OPTIMAL TRADING PROCESSES

Proceeding as before on the assumption that the trader's output, \(s\), is predetermined and equal to his measured money expenditure, \(d\), we may define an optimal stationary trading process as one that minimizes marketing costs, the latter being defined as the sum of four components: waiting costs, finance costs, storage costs, and trading costs. In order to establish explicit criteria for distinguishing optimal from nonoptimal processes, therefore, we must introduce appropriate functional relationships to link each of these cost components with the trader's decision variables, \(D\) and \(S\).

The waiting-cost function poses no problems. Given the information provided by the finance function (4), we may write

\[
WC = r[D + S + M] = r[D + S - G(D,S)] - rL
\]

where \(r\) denotes some appropriate rate of discount. In an economy that admits of costless borrowing and lending at a given market rate of interest, \(r\) might reasonably be interpreted as that market rate. More generally (i.e., in the absence of clearly specified external circumstances), \(r\) might be taken to represent an appropriate subjective rate of discount.
The finance-cost function reflects overdraft charges and the inflationary tax on money holdings. The interest rate charged on overdrafts is \(2r_m\), which we also suppose is the rate of interest paid on positive money holdings. We are ruling out any set-up costs associated with running an overdraft, and assuming that \(0 \leq 2r_m \leq r\). Assuming that the trader expects a rate of price inflation equal to \(2\pi\), we may write

\[
FC = (2\pi - 2r_m) \bar{M} = (\pi - r_m) [D+S-2G(D,S)] - 2(\pi - r_m)L
\]  

(6)

The definition of the storage-cost function is

\[
SC = 2\alpha \bar{D} + 2\beta \bar{S} + 2\gamma \bar{M} = \alpha D + \beta S + \gamma [D+S-2G(D,S)] - 2\gamma L
\]  

(7)

where \(2\alpha\), \(2\beta\), and \(2\gamma\) represent per unit storage costs on average holdings of each of the three commodities, and are all assumed to be non-negative. The coefficients \(\alpha\) and \(\beta\) are to be interpreted as reflecting not costs of physical wastage but rather costs of storage or costs associated with physical transport of commodities (i.e., costs of transacting that vary directly with quantities traded per transaction).

Having subsumed costs of transacting that vary directly with trans-
action quantities in the storage-cost coefficients \(\alpha\) and \(\beta\), we may plausibly suppose that trading costs proper are a function simply of the trading fre-
quencies \(s/D\) and \(s/S\). We may write the trading-cost function in general form as

\[
TC = T[D/s, S/s]
\]  

(8)

To lend minimal economic content to this relation we suppose that both of the first partial derivatives, \(T_1\) and \(T_2\), are negative, and that the function is homogeneous in both arguments; i.e., \(T[\lambda D/s, \lambda S/s] = \lambda^q T[D/s, S/s]\) for all
\( \lambda > 0 \), and some \( q < 0 \). The simple relation

\[
TC = \frac{as}{D} + \frac{bs}{S}
\]  

satisfies both of these conditions and is the mathematical equivalent of trading-cost functions in earlier literature.

In each of these cost functions the effect of the overdraft limit, \( L \), is just to add a constant term, which cannot affect the minimization problem. We shall therefore suppress all these constant terms, in which case we may treat all of the cost functions as being homogeneous in both arguments. Thus the total marketing-cost function of the trader is

\[
TMC = T[D/s, S/s] + \rho[D+S-G(D,S)] + \delta D + \varepsilon S
\]  

\[
\rho = r + 2\gamma + 2\pi - 2r_m
\]

\[
\delta = \alpha - \gamma - \pi + r_m
\]

\[
\varepsilon = \beta - \gamma - \pi + r_m
\]

Standard calculus techniques cannot be used directly to define extreme points of this function because of the discontinuities. Minima of the function can be adequately characterized for our purposes, however, in terms of a fairly straightforward, two-stage diagrammatic analysis.

Let the curves \( ab, cd, \) etc., in Figure 3 represent iso-trading cost (ITC) curves, the equations of which are \( T[D/s, S/s] = K \), where \( K \) is some constant. A necessary condition for given values of \( D \) and \( S \) to minimize \( TMC \) is that trading costs be minimal with respect to any given level of holding costs (finance costs plus storage costs plus waiting costs), since otherwise the transactor could choose another solution with the same holding costs but lower trading costs—that is, with lower total costs. Let us start accordingly by considering an arbitrary iso-holding cost (IHC) set—represented in Figure 3
Figure 3

The first-stage minimization problem
by points on or within the triangular area eʃ. The equation of this set (cf. equation (10), above) is

\[ H[D,S] = \rho[D+S - G(D,S)] + \delta D + \varepsilon S = J \]

where \( J \) is some constant. We note for future reference that the function \( H \) is homogeneous of degree one in both arguments. When \( D \) is an exact multiple or divisor of \( S \), the equation of the set is

\[ \rho \max(D,S) + \delta D + \varepsilon S = J \]

This equation is represented in Figure 3 by the two line segments eʃ and jf, the slopes of which are \( -\delta/(\rho + \varepsilon) \) and \( -(\delta + \rho)/\varepsilon \) respectively. The intersections of these two line segments with the integer rays \( S = 2D, S = D, S = D/2, \) etc., define points \( \ldots h, j, k, \ldots \) that lie on the upper boundary of the IHC set. All other elements of the set lie to the left of this boundary but on or to the right of the lower boundary (the line ef in Figure 3) which, by earlier results, is defined by the equation:

\[ \rho(D+S) + \delta D + \varepsilon S = J \]

Because of the discontinuities in the function \( G(D,S) \), the IHC set also has discontinuities. As we indicated earlier, the most important of these discontinuities occur at the integer rays. While points on these rays lie on the upper boundary of the IHC set any sequence of points approaching such a ray will approach the lower boundary, as indicated by points \( \ldots p, q, r, \ldots \).

As indicated above, a necessary condition for given values of \( D \) and \( S \) to yield a minimum of TMC is that they yield a minimum of TC for any given IHC set. This necessary condition is satisfied for the ITC curves and IHC set shown in Figure 3 only when \( D \) and \( S \) assume values that correspond to the point \( h \). Now consider other IHC sets corresponding to different values of \( J \).
Because both the holding-cost and trading-cost functions are homogeneous in \( D \) and \( S \), all points satisfying the required necessary condition for TMC to be minimal will lie on the ray (Oh in Figure 3) containing the point \( h \). It follows that the ratio of relative quantities traded, \( D/S \), is independent of the initial choice of an IHC set. This means that the solution to the decision problem can be decomposed into two stages. In the first stage, we imagine that the trader chooses an optimal trading ratio, \( w = D/S \), by minimizing trading costs on any given IHC set. In the second stage, we imagine that the trader arrives at optimal absolute values of \( D \) and \( S \) by choosing a level of sales that solves the standard calculus problem,

\[
\text{Minimize } T[wS/s, S/s] + H[wS, S] = S^q \cdot T[w/s, 1/s] + S \cdot H[w, 1];
\]

the solution to which is

\[
\hat{S} = \frac{1}{K(H[w, 1]/T[w/s, 1/s])^{q-1}}, \quad (12)
\]

with \( K = (-\frac{1}{q})^{q-1} \). Corresponding optimal values of \( \hat{D} \) and \( \hat{f} (= \hat{w} + \hat{L}) \) are:

\[
\hat{D} = w\hat{S} \quad \text{(13)}
\]

and

\[
\hat{f} = [(1+w)/2 - G(w, 1)] \cdot \hat{S} \quad \text{(14)}
\]

In the special case \( H = as/D + bs/S \) \( \hat{S} \) is given by the explicit square-root formula

\[
\hat{S} = \frac{\sqrt{[a/w+b]s}}{\sqrt{\delta w+e+\rho [w+1-G(w, 1)]}} \quad (15)
\]

so for this instance our results are formally analogous to those obtained in earlier work with inventory-theoretic models. That this formal analogy does not imply economic equivalence, however, is clear from the fact that in our
analysis the value of $\hat{S}$ depends in every case upon the value of the discontinuous function $G(\cdot,\cdot)$.

The significance of this observation may be brought out most easily by noting that for small variations in the parameters that determine the IHC sets and ITC curves, the value of the optimal transactions ratio, $w$, will not be affected because small variations in $w$ will be associated with relatively large variations in total marketing costs. There is a presumption, indeed, that the trader generally will choose a value of $w$ that corresponds to the slope of one of the integer rays in Figure 3. Thus, for large variations in the parameters that determine the form of the IHC sets and ITC curves, the value of $w$ may hop, as it were, from one integer-ray slope to another (roughly speaking, this will occur in situations where the ITC curves are relatively flat so that the ITC curve passing through a point of a given IHC set not on an integer ray will always pass to the left of points in the same IHC set that lie on immediately adjacent integer rays). But the rule suggested by presumption is not universal. To show this it is only necessary to construct an example where an ITC curve intersects an IHC set at two points on neighboring rays but also passes to the left of some point of the same IHC set between the two rays. In this case, the value of the optimum transactions ratio, $w$, will lie somewhere between the values of the slopes of the two integer rays. Such an example is defined for the special case $H = as/D + bs/S$ by supposing that $s=1$, $a=5/11$, $b=1$, $\rho=.01$, $\delta=.1$ and $\varepsilon=.1$. The two points at which an ITC curve and an IHC set intersect along neighboring integer rays are then given by

(i): $D = 100/21, S = 100/21$

(ii): $D = 200/32, S = 100/32$.

At (i) and (ii), trading costs are 6.72/22 and holding costs are 1. But at the point

(iii): $D = 300/54, S = 200/54$,
which lies on a ray between those defined by the points (i) and (ii), holding
costs are 1 and trading costs are 6.66/22. So it is sometimes optimal for the
trader to choose a trading point \((D, S)\) that does not lie on \(a\) integer ray.

IV. IMPLICATIONS

It would be injudicious to attach great significance to conclusions
derived from a model in which the trader is assumed to produce and sell just
one commodity and to purchase and consume just one other. By the same token,
however, it would be even more injudicious to attach general significance to
any proposition that is not implied by our model. In the immediately ensuing
discussion, we use this elementary criterion of restricted generality to assess
the robustness of various propositions that have been asserted (or conjectured)
by earlier writers.

A. Storage Costs and the Demand for Money. It has been argued by some
writers that the holding of positive money balances can be accounted for merely
in terms of the existence of transaction costs, and that physical characteristics such as durability, portability, divisibility, and so forth, are of
secondary importance. At least one author, however, has insisted that posi-
tive money balances will not be held unless storage costs on money are less
than those on other commodities. Our model indicates that neither of these
positions is generally defensible.

To show this, we first notice that a necessary condition for \(\bar{M} > 0\) in
our model is \(S \neq D\). Referring to Figure 3, this means that the optimal
trading point must lie off the ray \(D = S\). In order for this to be true, it
is necessary that the slope of the ITC curve in the neighborhood of the ray
\(D = S\) be negative and greater (in absolute value) than the negative slope of
the line segment \(jf\), or negative and less (in absolute value) than the nega-
tive slope of the line segment \(ej\). That is,
\[
0 < \frac{T_2}{T_1} < \frac{\delta}{\epsilon + \rho}, \quad \text{or} \\
0 < \frac{(\delta + \rho)}{\epsilon} < \frac{T_2}{T_1}
\]

Of course the trader might choose to hold money just because of the interest that it bears, or because he expects a negative rate of inflation. Since these motives are irrelevant to the present issue we may rule them out by assuming that \( \pi = r_m = 0 \), in which case (16) reduces to

\[
0 < \frac{T_2}{T_1} < \frac{(\alpha - \gamma)}{(\beta + r + \gamma)}, \quad \text{or} \\
0 < \frac{(\alpha + r + \gamma)}{(\beta - \gamma)} < \frac{T_2}{T_1}.
\]

Obviously neither of these conditions can be satisfied unless either \( \alpha > \gamma \) or \( \beta > \gamma \). Thus it would seem that one of the other commodities must be more costly to store than money if positive money balances are to be held. However, in interpreting this conclusion of our formal model we must remember that the question of phasing has been assumed away. While a sufficiently large storage cost on money will clearly induce a trader to set the frequency of purchases equal to the frequency of sales, a more general analysis might still allow for a positive demand for money in this situation because there might be a prohibitive cost to exact synchronization. Thus a trader would still hold money during the interval between each sale and the following purchase.

Two conclusions follow directly from this line of reasoning. First, no explanation of the demand for money and the existence of monetary exchange arrangements can be adequate that does not take explicit account of storage, transportation, and other costs that depend directly upon the physical quantities traded,\(^{16}\) because without such an account only the "float" can be explained. Second, we may say that the existence of these costs is generally not necessary in order for the demand for money to be positive.

**B. The Demand for Money and Inflation.** Modern studies of hyper-inflation indicate that even when inflation rates—and so costs of holding
money balances--rise to extreme levels, individuals are extremely reluctant to forego trade in organized markets that require them to use conventional media of exchange. This kind of behavior is consistent with our model if there are in fact bunching costs that prevent the perfect synchronization of transactions. Otherwise our model can be shown to imply that the demand for money goes to zero when rates of (expected) inflation exceed prevailing rates of storage cost on non-money commodities; i.e., beyond some critical point the inflationary process will explode for there will be no finite price level that will equate the supply and demand for money. This result, considered in conjunction with the empirical evidence mentioned above, casts serious doubt on the general validity of the conventional assumption that the question of phasing may be ignored.

If we were to take explicit account of the phasing question, the present approach would seem to suggest that during hyperinflations the frequencies of purchase and sale would tend toward equality, and the interval between a sale and the subsequent purchase would shorten considerably. That is, the demand for money would reduce to the float, and the size of the float would tend to diminish. This would appear to be consistent with the recorded experience of hyperinflations, during which the frequency of wage and salary payments, which under normal circumstances is much less than the typical household's frequency of shopping trips, tends to rise sharply, and approaches the frequency of shopping trips, as workers rush to spend each payment before the next one is received.

C. Comparative Statics Problems. One serious consequence of the discontinuities in the G-function is that they make it impossible to predict the outcome of many comparative statics experiments with any degree of generality. In the face of these difficulties it is natural to try imposing superficially plausible a priori restrictions on the finance function that might eliminate
the discontinuities. That models constructed on this basis are, at best, approximations to some "true" model is generally recognized. What is perhaps not so generally recognized is that the solution-values generated by such models may closely approximate the corresponding solution-values of an underlylng "true" model and yet the comparative静态s results derived from these solution-values might be qualitatively misleading or wrong.

By way of illustration, suppose that we decided to ignore the inherent fuzziness of our finance function and to replace it by an approximation—specifically, the approximation that projects all points of the "true" relation onto its lower boundary. This procedure could be defended on the grounds that points actually chosen by a trader are likely in any case to lie on the lower boundary and, even if that were not so, would always lie "close" to it. Suppose that we also imposed two additional restrictions, namely, that \( S > D \) (after all, salary checks usually are larger than single purchases of commodities) and that the trading cost function is given by (9). On these assumptions the total marketing cost function would be

\[
TMC = \frac{as}{D} + \frac{bs}{S} + \delta D + (\varepsilon \rho)S.
\]

Optimal values of \( D \) and \( S \) would then be given by the equations

\[
D = \sqrt{\frac{as}{\alpha \gamma - \pi + r_m}} \quad \text{and} \quad S = \sqrt{\frac{bs}{\rho + \beta \gamma + \pi - r_m}}
\]

which imply (among other things) that \( \frac{\partial D}{\partial r_m} < 0 \).

This is the result obtained by Feige and Parkin [1971] using a similar approximate approach. In the "true" model that underlies our approximation, however, it is clear that a small increase in \( r_m \) might produce an increase in \( D \). In particular, suppose the transactor is initially on an integer ray. As indicated earlier, it is possible that a small change in \( r_m \) may have no effect on the relative trading ratio, \( D/S \), which is to say that \( D \) and \( S \) will
change in the same direction. In fact, the analysis of the next section implies that they must both increase, for otherwise there could be no reduction in transaction costs. But this is just one possibility. The "true" model also might produce exactly the same result as the approximate model. This would happen if, in Figure 3, the line ej was tangent to an ITC curve along an integer ray both before and after the change in \( r_m \).

The economic interpretation of this ambiguity is straightforward. If we ignore solutions off the integer rays, it appears that when \( r_m \) is increased the transactor can augment money holdings by increasing the absolute size and reducing the relative size of purchases. The relative size effect will increase \( \bar{M} \) because a transactor who sells every week and buys n times a week will have average money holdings of \( \bar{M} = (1-1/n)\bar{S} \), which may be increased either by increasing \( \bar{S} \) or by increasing \( n \). In the approximate solution it turns out that the relative size effect dominates. In the "true" solution, there might be no relative size effect because a small change in relative size would result in too large a discontinuous increase in \( \bar{M} \).

Another comparative statics result that holds for our approximate model (and for other similar models) but appears to break down in the "true" model is the proposition that \( \bar{M}/\partial r < 0 \). An increase in \( r \) will induce a reduction in total inventory holdings in the "true" model, but holdings of any single inventory may increase. Similar ambiguities arise in connection with other comparative statics results that have been obtained in earlier literature.\(^{18}\) However, one classic result continues to hold for our model in a form that is even more general than the one derived from earlier inventory-theoretic models—provided that the trading cost and holding cost functions are homogeneous. That is the proposition that the income elasticity of demand for money and other inventories is one half.\(^{19}\) This follows because a change in \( s \) will not
affect the IHC sets, nor will it affect the slopes of the ITC curves. Thus it will not affect \( w \), the relative size of purchases, so the income elasticities may be derived from (12) \( \sim \) (14), where \( w \) is held constant. If the T-function is homogeneous of degree \( q \) then

\[
\hat{S} = K \left( \frac{H[w,1]}{T[w,1]} \right)^{\frac{1}{q-1}} \cdot s^{\frac{q}{q-1}}
\]

It follows that the income-elasticity of \( \hat{S}, \hat{D} \), and \( \hat{F} \) is \( \frac{q}{q-1} \). In the special case where the trading cost function is given by (9), this elasticity is one half.

D. The Optimum Quantity of Money. It is well known that Pareto optimality requires \( 2r_m \) to be set equal to \( r+2\pi. \) \(^{20}\) It is generally presumed that in existing monetary systems \( r_m \) is less than this amount. However, there does not appear to be any agreement about precisely what would happen, in a qualitative sense, if \( r_m \) were increased. \(^{21}\) On the basis of the preceding analysis it is clear that the effects of changes in \( r_m \) on the values of \( \hat{D} \) and \( \hat{S} \) cannot generally be predicted. However, our analysis also indicates that as long as we start with \( \hat{F} \neq 0 \), then an increase in \( r_m \) will be associated with both an increase in average holdings of money balances and a decrease in trading costs.

The first conclusion follows from a simple argument of the kind first used by Samuelson in the context of production and consumption theories. \(^{22}\) Suppose that, when \( r_m = r_m^o \), the transactor chooses \( S^o, D^o, \bar{M}^o \), and when \( r_m = r_m' \), ceteris paribus, the transactor chooses \( S', D', \bar{M}' \). In the first case the transactor could have chosen the second solution. The fact that he did not implies that when \( r_m = r_m^o \) the total cost of the second solution is greater than or equal to the total cost of the first solution. If we combine this fact with
the analogous fact in the second case, it is then a matter of straightforward
calculation to show that $(r_m' - r_m^O)(\bar{M}' - \bar{M}^O) \geq 0$, which implies that average
money balances must increase if they change at all. Furthermore, they will
indeed change unless $\bar{M}^O = \bar{M}^L = -L$. 23

The second conclusion follows from the fact that total trading costs
associated with an optimal solution to the second-stage of the trader's de-
cision problem for any given value of $r_m$ are constantly proportional to total
trading costs associated with optimal solutions for other values of $r_m$. More
precisely, it follows from (12) that

$$\hat{s}^q \cdot T[w/s, 1/s] = K \cdot \hat{s} \cdot H[w, 1].$$

Therefore,

$$T[\hat{D}/s, \hat{s}/s] = KH[\hat{D}, \hat{s}],$$

which is the symbolic equivalent of the conclusion stated above. Since total
marketing costs must go down (provided again that $\hat{F} \neq 0$), when $r_m$ increases,
it follows immediately that total holding costs and total trading costs must
change (if at all) in the same direction.

The same conclusion has been reached by other writers who have put money
into utility or production functions as a proxy for savings in transaction time. 24
However, there is no sense in which the present analysis can be seen as a special
case of that type of analysis, because in our analysis it would be impossible to
define the marginal utility or marginal product of money. Money holdings may be
regarded as an ingredient input to the transactions technology. In order for an
extra unit of money to confer any benefit upon the holder (other than interest
return) it must be accompanied by a change in D or S. 25

Does the preceding argument entitle us to assert that payment of interest
on money would normally imply a reduction in real costs of trading activity?
It might, if we had faith in the assumption (critical for the entire analysis) that the trading cost function $H$ is homogeneous. In truth, we have no grounds for this assumption other than considerations of analytical convenience. Suppose that it were false. Suppose, further, that the technology of exchange were such that the "expansion path" corresponding to the locus of possible optimal solutions to the trader's decision problem was backward-bending. Then our first result (that $\bar{M}$ would never decrease in response to an increase in $r_m$) would still hold, since the proof of this result does not depend on the transactions technology; but the second result might well be reversed. That is to say, money might be an "inferior factor" in the production of transaction services.  

V. EXTENSIONS

The model underlying the preceding analysis is too special to serve as anything more than a propaedeutic to a full study of the inventory demand for money. As it stands, the model is valuable mainly as a supplement to earlier work and as a guide to further research; it does not deserve to be taken seriously as even a first approximation to the description of observable patterns of transactor behavior. By focusing on fundamentals in a precise way, however, our model brings out clearly certain characteristics that a more general model should have in order to be taken seriously. That is the theme to which we address ourselves in the remainder of this paper.

One question that naturally arises in connection with any few-commodity model is whether qualitative results established for the special case carry over to many-commodity generalizations. In the present instance, for example, it is conceivable that an extension of our model to deal explicitly with $m$ production goods and $n$ consumption goods might yield a somewhat smoother demand function for money than is obtained for the $1 \times 1$ case. It is also
conceivable that other propositions (e.g., assertions about income elasticities of demands for inventories) might fail to hold in a more general model.

Unfortunately, a general multi-commodity formulation of our analysis, although possible in principle, does not appear to be feasible in practice—or if it is, we have been unable to discover a way of carrying it through. The main difficulty lies in the fact that with more than two non-money commodities, the trader's required average money balance depends not only upon the size and frequency of transactions but also upon their phasing.

As an example, suppose that there are two consumption goods and one production good, that the two flow rates of consumption are equal, and that the size of each type of transaction is equal to one week's income. There are at least two ways to phase these transactions. One is to make one sale at the end of each week, and to purchase one commodity at the end of each even week and the other commodity at the end of each odd week. In this case the average money holding would be zero because each sale would coincide with a purchase of equal value. Another way would be to make two purchases and one sale at the end of each even week and one sale at the end of each odd week. This would imply average money holdings of one-half of a week's income, because money holdings would be equal to zero during each odd week and equal to one week's income during each even week. In this example, it is evident that the minimum required average money balance would be obtained by choosing the first rather than the second phasing arrangement. In more complex cases, however, the problem of optimal phasing has no such simple and direct solution because, in general, no mere change in phasing will achieve perfect synchronization of purchases and sales.

Though we have not been able to arrive at an explicit characterization of the finance function for the m x n case, we are tentatively persuaded
(on the basis of casual analytical investigation) that our understanding of
the inventory demand for money would not be much advanced had we been suc-
cessful. Specifically, it appears that if the trading cost and holding cost
functions of an mxn model satisfy appropriate homogeneity conditions (and we
can see no reason why mere number of commodities traded should be of any
significance in this regard), then all qualitative propositions that hold
(or fail to hold) in the 1x1 case should have their counterpart in the mxn
case. In particular, there is no reason to believe that the demand function
for money would be smoother in the mxn than in the 1x1 model; the discon-
tinuities that occur in the latter model clearly derive not from the assump-
tion that only two non-money goods are traded but rather from the assumption
that trading dates are evenly spaced and adhered to rigidly in any stationary
trading process. We shall not dwell further on these matters here, however,
because it is evident on other grounds that even the most detailed knowledge
of the properties of an mxn model would add little of economic interest to
what can be said on the basis of the 1x1 model.

The crucial weakness of the present model--as of inventory-theoretic
models generally--lies in the implicit assumption that cash payments and re-
ceipts are rigidly linked with commodity purchases and sales. This clearly
is not true for more than a small minority of market transactions in any econ-
omy of record. In practice, most transactors have access to a variety of
credit facilities--including established lines of trade credit--that are much
more flexible than the simple overdraft arrangement that we have postulated.
These facilities permit them to acquire goods in accordance with one timing
pattern and to pay for them according to another. The effect of such arrange-
ments--in language appropriate to our model--is to make the lower boundary
of the finance function a soft mattress rather than a hard floor. The
existence of such lines of credit incidentally, introduces a stochastic element into the typical trader's cash flow since sales will frequently be separated from receipts by a significant and variable time interval. Transactors in the real world generally have access also to profitable lending opportunities, the effect of which is to take the sting out of trading processes that entail average holdings of money balances near the upper boundary of the money requirements function. Of course, lending opportunities cannot be used effectively by small-scale holders of transactions balances; investment fees would more than wipe out earnings. Large-scale businesses can earn substantial sums by temporary investment of cash surpluses, however, and many firms have been quick to exploit such opportunities in recent years.

To introduce these and related considerations into our theory would be extremely difficult—but probably most worthwhile. It is an open question, of course, how such extensions would affect specific conclusions reached in this paper, since that would depend critically on what assumptions were made about set-up costs on borrowing and lending and about the effect of inventory management costs and profits upon income and expenditure flows. It is evident without detailed argument, however, that the rigidities and discontinuities that play such a critical role in our simple model would largely disappear in the more general case. We conjecture, indeed, that the development of borrowing, lending and trade-credit facilities in the real world has occurred, in part, in response to attempts by transactors to circumvent the difficult cash management problems that obviously would arise in a world that conformed at all closely to the rigid specifications of the model presented in this paper.
Mathematical Appendix

This appendix contains the proofs of two theorems. The first theorem establishes formula (2) in the text. The theorem is a generalization of an earlier theorem that was proved by Georges Monette of the University of Western Ontario. Monette's theorem, which played a crucial role in the preparation of the present paper, established formula (2) for the special case where D and S are integers.

Notation:

1) \( G(D,S) = \begin{cases} 
\text{the largest common divisor of D and S, when} \\
\text{D/S is rational (see fn. 10 above),} \\
0, \text{ when D/S is irrational.} 
\end{cases} \)

2) \( [x] \) denotes the largest integer no greater than the real number \( x \).

3) \( V(t) = \frac{tS}{S} + 1 S - \frac{tS}{D} + 1 \cdot D \), the virtual time path of money (where we have set \( t_o = 0 \)).

Theorem 1: For any \( D > 0 \) and \( S > 0 \), \( \inf_{[t \geq 0]} V(t) = -D + G(D,S) \).

Proof: Define \( \phi(t) = V(t) + D = \frac{tS}{S} + 1 S - \frac{tS}{D} + 1 \cdot D \). We shall prove that

\( \inf_{[t \geq 0]} \phi(t) = G(D,S). \) By inspection of \( \phi(t) \) it is clear that

\( \phi(t) \geq 0. \) It is also clear that when \( D/S \) is rational, then

\( \phi(t) \) is divisible by \( G(D,S); \) that is, \( \phi(t) = k G(D,S) \) for some positive integer \( k. \) Therefore,

\( \phi(t) \geq G(D,S) \) for all \( t, D, S. \) \( \) (A.1)
(a) Suppose D/S is rational. Then, by a simple generalization of a well known result* we know that there exist positive integers u and v such that

\[ uS - vD = G(D,S). \]

Define \( t' = \frac{vD}{s} \). Then:

\[ t's = vD = uS - G(D,S) \]

\[ \frac{t's}{s} = u - G(D/s,1) \Rightarrow u - 1 \]

\[ \left\lfloor \frac{t's}{s} \right\rfloor = u - 1. \]

Thus we have:

\[ \phi(t') = \left\{ \left\lfloor \frac{t's}{s} \right\rfloor + 1 \right\} \cdot S - \left\lfloor \frac{t's}{D} \right\rfloor \cdot D \]

\[ = uS - vD \]

\[ = G(D,S) \quad \text{(A.2)} \]

In this case the theorem follows from (A.1) and (A.2).

(b) Suppose D/S is irrational. Define \( \xi = D/S \). There exist two sequences of positive integers, \( \{\alpha^n\} \) and \( \{\beta^n\} \), such that**

\[ \beta^n - \alpha^n \cdot \xi > 0, \quad \text{and} \]

\[ \beta^n - \alpha^n \cdot \xi \to 0. \quad \text{(A.3)} \]

\[ \beta^n - \alpha^n \cdot \xi \to 0. \quad \text{(A.4)} \]

It follows from (A.3) and (A.4) that:

\[ [\alpha^n \cdot \xi + 1 - \beta^n] \to 0 \quad \text{(A.5)} \]

Consider the sequence of dates \( \{t^n\} = \{t^n\} = \{\frac{\alpha^n \cdot D}{s}\} \).

* Niven and Zuckerman (1972), p. 4.

** Niven and Zuckerman (1972), pp. 134-139.
\[ \phi(t^n) = \left( \frac{\alpha^n \cdot D}{S} + 1 \right) \cdot S - \alpha^n \cdot D \]
\[ = \left[ (\alpha^n \cdot 5 + 1) - \alpha^n \cdot 5 \right] \cdot S \]
\[ = [\alpha^n \cdot 5 + 1 - \beta^n] \cdot S = (\alpha^n \cdot 5 - \beta^n) \cdot S \]

Therefore, from (A.4) and (A.5):

\[ \phi(t^n) \to 0 = G(D,S) \]  \hspace{1cm} (A.6)  

The theorem follows in this case from (A.1) and (A.6).

Q.E.D.

The second theorem establishes (P.4) in the text.

Theorem 2: Suppose D/S is rational. Then for any sequence \( \{D^q, S^q\} \)

such that (a) \( D^q/S^q \neq D/S \) for all \( q \); (b) \( D^q > 0 \) and

\( S^q > 0 \) for all \( q \); and (c) \( \lim_{q \to \infty} (D^q, S^q) = (D, S) \), then

\[ \lim_{q \to \infty} G(D^q, S^q) = 0. \]

Proof: Because D/S is rational, there exist positive integers \( m \) and \( n \)

such that \( mD = nS \). Therefore, from (c),

\[ \lim_{q \to \infty} mD^q = mD = nS = \lim_{q \to \infty} nS^q \]  \hspace{1cm} (d)  

Take any \( q \). If \( D^q/S^q \) is rational then, from (a) and (b),

\[ \frac{mD^q}{G(D^q, S^q)} \text{ and } \frac{nS^q}{G(D^q, S^q)} \text{ are distinct integers, and } G(D^q, S^q) > 0. \]

Therefore, \[ \left| \frac{mD^q}{G(D^q, S^q)} - \frac{nS^q}{G(D^q, S^q)} \right| \geq 1; \text{ that is,} \]

\[ |mD^q - nS^q| \geq G(D^q, S^q) \geq 0 \]  \hspace{1cm} (e)  

Obviously the inequality (e) will also be satisfied if \( D^q/S^q \) is irrational. Taking limits in (e), and noting that, from (d),

\[ \lim_{q \to \infty} |mD^q - nS^q| = 0, \text{ we get } \lim_{q \to \infty} G(D^q, S^q) = 0. \]

Q.E.D.
References


Footnotes

* The authors are, respectively, Assistant Professor of Economics at The University of Western Ontario, and Professor of Economics at the University of California, Los Angeles. We are indebted to J. S. Fried and J. M. Ostroy for suggestions and critical comments on earlier versions of the paper. We have also benefited from reading the results of investigations on a closely related problem as reported in a privately circulated memorandum from D. W. Bushaw (see footnote 7, below).


2 See Clower (1967) and Hahn (1965).

3 This point has been made most forcibly by Ostroy (1973). See also Clower (1967) and Hirshleifer (1973).

4 This approach is outlined by Clower (1970), and has been adopted in one form or another by Johnson (1970), Perlman (1971), Feige and Parkin (1971), Fried (1973) and Barro and Santomero (1973).


6 See Howitt (1973) for a formal model of such an economy.

7 The problem was studied in some detail by D. W. Bushaw and five other mathematicians during a summer institute in applied mathematics at Washington State University in 1972. Our judgment about the difficulty of the problem is based on the paucity of firm results obtained by this group (which included some leading specialists in dynamical polysystems).

8 This condition makes sense only if trade-related costs are implicit (e.g., foregone leisure) or are financed from resources that do not depend directly on quantities traded. The model could be rewritten to treat trade-related costs as explicit deductions from sales receipts and explicit additions to purchase costs, but that would just add notational complications without altering economic substance, so we follow the simpler procedure of supposing that such costs are met "outside" the model.

9 This makes sense when we are talking about bunching a purchase closely with a sale. However if the two transactions are both purchases or both sales, the costs are more likely to decrease with closer bunching. An example would be purchases of different types of groceries in one shopping trip. This alternative assumption is made by Fried (1973) and Reinhardt (1973).
10. This requires that we extend the usual definition of largest common divisor to apply to numbers that are not both integers. This can be done by defining it as the largest real number by which the two arguments may be divided so that the resulting numbers are both integers.

11. See Niven and Zuckerman (1972), pp. 4-6.

12. We use the term trading costs rather than the more usual term transaction costs because our definition of the former concept is more restrictive than the standard definition of the latter (which typically includes certain elements that we regard as storage costs).


15. Of course this is not a sufficient condition in our model because of the overdraft privileges. In our model the question should be reinterpreted as being whether or not the demand for finance, \( \bar{M} + L \), will be positive.


18. Some of these results may continue to hold in a restrictive sense. If the approximation of solution-values is close enough, then for large enough changes in parameters the changes in the "true" solution-values might be so large that the close approximation must change in the same direction as the "true" solution-values in order for these approximations to remain close.

19. See Baumol (1952).

20. See, for example, Samuelson (1968). For a recent criticism of certain aspects of the usual optimality argument, see Gramm (1974).


23. To see this, suppose that \( \bar{M}_0 \neq -L \), but that there is no change in \( \bar{M} \). It follows that there will be no change in \( w \); for if a change in \( w \) is required for \( r^m \), it would also have been required for \( r^0_m \). Thus the changes in \( \bar{D}, \bar{S}, \) and \( \bar{F} \) can be derived from (12) \( \sim \) (14) with \( w \) unchanged. But in this case we get
\[
\frac{d\delta}{dr_m} = \frac{\partial}{\partial r_m} \left\{ \frac{(r + 2\gamma + 2\pi)(w + 1 - G(w, 1)) + (\alpha - \gamma - \pi)w + \beta - \gamma - \pi - r_m \tilde{M}}{T[w/s, 1/s]} \right\} \cdot K^{1/\pi - 1}
\]

which can only be zero if \( \tilde{M} = 0 \).

24 Cf. Samuelson (1968). The same technique has been used, but in a more roundabout way, by Saving (1971).

25 Fischer (1974) has argued that in a similar context the marginal product or utility of money holdings might be defined. However, his argument is based upon the Baumol-Tobin analysis where the only transactions are bond transactions, in which case knowledge of the average holdings of money balances is sufficient to determine the pattern of transactions. Thus a change in \( \tilde{M} \) has a determinate necessary effect on the pattern of transactions in his analysis but not in ours. The discontinuities uncovered in the present analysis pose another difficulty in defining the marginal product or utility of money.

26 Lack of homogeneity would also add one more source of ambiguity to the comparative statics results mentioned in the previous section.