Revised October 1977

THE SUPPLY OF DEALER SERVICES
IN SECURITIES MARKETS

by

HANS R. STOLL

Working Paper No. 2-78

RODNEY L. WHITE CENTER
FOR FINANCIAL RESEARCH

University of Pennsylvania

The Wharton School

Philadelphia, Pa. 19104

This is a revision of an earlier paper written at the University of Pennsylvania. I benefited greatly from the comments of my colleagues at the University of Chicago while a visitor there in 1975-76, from the comments of my colleagues at Ecole Superieure des Sciences Economiques et Commerciales while a visitor there in 1976-77, as well as from the comments of my colleagues at the University of Pennsylvania. Special thanks are due Marshall Blume, Bernard Dumas, Thomas Epps, Robert Hamada, Jonathan Ingersoll, Robert Litzenberger, Ron Masulis, Merton Miller, James Morris, Roland Portait, Randolph Westerfield and Joseph Williams. All are absolved from the remaining errors. The financial support of the University of Pennsylvania Law School Center for the Study of Financial Institutions is gratefully acknowledged.

THE SUPPLY OF DEALER SERVICES IN SECURITIES MARKETS

There has been much discussion of a policy nature about the function of dealers in equity securities markets, the efficiency of different methods of providing dealer services and the regulatory constraints under which dealers should operate. Some empirical work has been carried out to determine what factors underlie dealer costs and to assess the efficiency of different market organizations and regulatory constraints, but that work has not been based on a very explicit theoretical foundation [see Demsetz (1968), Institutional Investor Study, Ch. 12 (1971), Tinic (1972), Tinic and West (1972), and Benston and Hagerman (1974)]. The purpose of this paper is to develop a more explicit and rigorous model of the individual dealer and to discuss the implications for the cost of trading of different market organizations of dealers. It is hoped this model will provide a better framework for empirical work and for discussion of the policy issues involved. The paper is restricted to the supply side. For steps in the direction of specifying the demand for dealer services-see Copeland (1976) and Epps (1976).

Dealers facilitate trading by investors because they are willing to trade for their own account as principals when investors' agents (or investors acting for themselves) cannot immediately find other investors with whom to trade.[1] Following Demsetz (1968) one can, therefore, think of dealers as providing the service of immediate trading or immediacy. The cost of immediacy developed in this paper is the sum of: (1) holding costs, the price risk and opportunity cost of holding securities; (2) order costs, the costs of arranging trades, recording and clearing a transaction; and (3) information costs which arise if investors trade on the basis of superior information.

Dealers are to be distinguished from brokers who act strictly as agents for investors and do not assume risk. This paper emphasizes holding costs, which have also been the emphasis of earlier studies on dealers. Order costs, which are incurred both by dealers and brokers, are included in the analysis but do not receive extensive development. Information costs facing dealers are also modeled, but see Jaffe and Winkler (1976) for a more extended treatment of that issue. Dealers exist because they provide immediacy more cheaply than investors could provide it for themselves (by placing limit orders with brokers and borrowing or lending until the limit order is executed, for example). Not everyone is a dealer because there are fixed costs (office, phone, etc) to the dealer business.

Section I presents an intuitive approach to holding costs. Section II of the paper discusses the principal assumptions underlying the analysis. Section III develops the holding cost function in a one period context. Implications of the model for determinants of bid-ask spreads and for the role of diversification by dealers are discussed in Section IV. Order costs and information costs are discussed in Section V. In Section VI certain policy issues related to the industrial organization of dealers are discussed in the context of the model. In Section VII modifications are made to put the model in a multiperiod context. The conclusions are summarized in Section VIII.

I. A Familiar Picture

The dealer can be viewed as any investor who has a desired portfolio (his investment account) based on the opportunities he sees and on his preferences. Supplying immediacy to other investors means moving away from this desired portfolio in order to accommodate the desires of investors to buy or sell a stock in which the dealer specializes. As a result the dealer assumes unnecessary risk and moves to a level of risk and return which may be inconsistent with his personal preferences.

These points are best illustrated by considering Figure 1. Line R_fE is the dealer's efficient frontier, which represents the possible combinations of his efficient portfolio of risky assets (point E) and the risk free asset (yielding R_f). Assume the dealer's desired position--his investment account--is point N. By taking nonoptimal portfolio positions the dealer moves away from N and indifference curve U_0 to a lower indifference curve such as U_1. The dealer is forced off the efficient frontier because it is assumed that he cannot initiate trades in his efficient portfolio of stocks, E. The portfolio acquired in the process of acting as a dealer is termed the trading account. Long or short positions in his trading account may de-diversify the dealer's total portfolio and cause the frontier of his new portfolio set (trading plus investment account) to be described by a line such as ANB, the position of which depends on the risk characteristics of the trading account.

A movement upward along NA means the dealer has an undiversified long position financed by borrowing at R_f. At point A he has, for example, borrowed 100% of his wealth. This imposes a total percentage holding cost of g on him. In other words, his customers must pay him g per cent on his current wealth in order to keep the dealer at his initial indifference curve. Part of the cost is due to the de-diversification caused by dealing in few stocks and part is due to the assumption of a level of risk not consistent with the dealer's preferences. Even if a dealer were to trade a fully diversified portfolio and thereby maintain R_fE as his frontier, he incurs costs because he moves to a lower indifference curve. A movement along NB means the dealer has an undiversified short position in the trading account.

Once the dealer is at a nonoptimal point like A, the cost of another transaction is the difference between the total percentage cost at A and at the new position resulting from the transaction. The cost may be negative if the dealer decreases his position and trades a different stock so as to increase diversification and reduce risk. The case of a decrease in position accomplished by selling another stock is illustrated in Figure 1 by the move from A to A_1. Since $g_1 < g$, the cost to the dealer of the transactions is negative, i.e., he is willing to pay customers to take him from A to A_1. The cost would be positive were the transactions to increase the dealer's position and reduce diversification. If the dealer specializes in one stock, his movement is along a curve like ANB. If he makes a market in many stocks, many paths are possible.

The task now is to make more explicit the holding cost incurred by a dealer and its relation to the size of the transaction and other variables. Before doing so the assumptions underlying the model are set out.

II. Assumptions

1. Dealer inventory positions acquired in the process of providing immediacy are financed solely at the risk free rate of interest, R_f. Thus dealer purchases of shares are financed solely by borrowing at R_f, and the proceeds of short sales are invested solely in the risk free asset. The dealer's personal wealth (his investment account) and the position in the trading account serve as collateral for the borrowing of cash or of shares.

Ruled out under this assumption is the possibility that dealers can acquire funds by selling other stocks or that they use the proceeds of short sales to buy other stocks. To permit this would simply transfer costs of immediacy to another dealer.

However, by appropriate changes in his bid-ask quotations the dealer encourages transactions by the public that will rebalance his portfolio. In other words, the dealer acts passively setting prices and letting the public choose which stocks it will purchase from him and which stocks it will sell to him. Portfolio adjustments arising from his dealer activity are therefore restricted to those stocks in which he makes a market, i.e. his trading account. The assumption of constant R_f could be relaxed at the cost of complicating the model, and the possibility is discussed below.[2]

2. The dealer is assumed to have a utility function over terminal wealth. Since the vast majority of dealers are proprietorships, partnerships or closely held corporations, ascribing a utility function to the dealer is realistic.[3] Alternatively one can assume the utility function belongs to the owners of a dealer firm and that the owners cannot offset on personnel account the positions taken by the firm (because they are unaware of the positions or because it is too costly for them).

3. The dealer makes estimates of the "true" price and "true" rates of return that would exist in the absence of transaction costs. This "true" price is the discounted value of the dealer's expected equilibrium price one period hence. This estimate is derived from the fundamental characteristics of the stock and need not be the same as the estimates of other dealers or investors. All rates of return in the paper are "true" rates of return. The analysis is a partial equilibrium analysis of the dealer industry and the determinants of "true" equilibrium prices is for example not treated.

4. The dealer makes one transaction per trading interval during which the stock's price does not change. Prices may change between trading intervals. In a one period world, the dealer buys or sells shares in the first trading interval and becomes subject to one period of uncertainty. The period is assumed to be very short, and certain approximations to be specified later may be justified. The world ends in the second trading interval when the dealer's inventory is liquidated at the equilibrium price of the second trading interval.

Certain other assumptions or lesser importance are detailed in the development of the model. These involve constraints on the utility function and certain approximations.

III. The Holding Cost Function

The dealer trades a stock if the dollar compensation paid him is enough to offset the loss of utility caused by deviating from his initial portfolio. In other words he will require that expected utility of terminal wealth of the initial and new portfolios be the same :

$$EU(\tilde{W}^*) = EU(\tilde{W}) \qquad (1)$$

where

$\tilde{W}^*$ = terminal wealth of the initial portfolio.

$\tilde{W}$ = terminal wealth of the new portfolio after the transaction.

$\sim$ indicates random variable.

The initial portfolio is a combination of the dealer's investment account (represented by point N in Figure 1) and his initial trading account. Thus

$$\tilde{W}^* = W_0\left[1 + k\tilde{R}_e + \frac{Q_p}{W_0}\tilde{R}_p + \left(1 - k - \frac{Q_p}{W_0}\right)R_f\right] , \qquad (2)$$

where

W_0 = initial wealth.

k = optimal fraction of the dealer's wealth invested in portfolio E, a constant because of assumption 1.

$\tilde{R}_e$ = return on portfolio E--the optimal efficient portfolio of risky assets. (Under homogeneous expectations this would be the so-called market portfolio that includes all risky assets.)

Q_p = "true" dollar value of stocks in trading account. Although one period remains, the dealer is allowed to enter the trading interval with a non-zero trading account, acquired in prior periods. If $Q_p = 0$, the initial portfolio is the desired portfolio -- point N in Figure 1. $Q_p \lessgtr 0$ according as the dealer as long or short position in the trading account.

$\tilde{R}_p$ = rate of return on the trading account.

R_f = risk free rate.

To simplify exposition, (2) may also be written as follows :

$$\tilde{W}^* = W_0(1 + \tilde{R}^*) \qquad (3)$$

where $\tilde{R}^*$ is the rate of return on the initial portfolio and is defined by the last three terms inside the brackets of (2).

Terminal wealth under the new portfolio is given by :

$$\tilde{W} = W_0(1 + \tilde{R}^*) + (1 + \tilde{R}_i)Q_i - (1 + R_f)(Q_i - C_i) \qquad (4)$$

where

Q_i = "true" dollar value of the transaction in stock i, the stock in which immediacy is being provided. Negative values indicate a sale; positive values, a purchase.

$\tilde{R}_i$ = rate of return on stock i.

C_i = present dollar cost to the dealer of trading the amount Q_i. C_i is positive or negative according as the transaction in stock i raises or lowers the costs of holding the inventory Q_p.

The dealer's cost, C_i, is incorporated as in (4) because, under current institutional arrangements, C_i, is not paid explicitly to the dealer at the time he provides immediacy. Instead the dealer trades at the bid or ask price different from the "true" price of the stock. Thus he need borrow only Q_i-C_i to finance a purchase the "true" value of which is Q_i. On a short sale he earns interest on Q_i+C_i although the present "true" value of the short sale is Q_i.[4]

Approximating (1) by expanding each side in a Taylor series around the respective means and dropping terms of order higher than two yields[5]:

$$E\left[U(\bar{W}^*) + U'(\bar{W}^*)\ (W^* - \bar{W}^*) + 1/2U''(\bar{W}^*)(W^* - \bar{W}^*)^2\right]$$
$$= E\left[U(\bar{W}) + U'(\bar{W})(W - \bar{W}) + 1/2U''(\bar{W})(W-\bar{W})^2\right] \tag{5}$$

where the bar (-) over a variable indicates expected value and where tildes have been dropped when the meaning is clear. Writing W and W^* in terms of initial wealth and rates of return and taking expectations yields :

$$U(\bar{W}^*) + 0 + 1/2U''(\bar{W}^*)W_0^2\ \sigma_*^2 = U(\bar{W}) + 0$$
$$+ \tfrac{1}{2}U''(\bar{W})\ \left[W_0^2\ \sigma_*^2 + Q_i^2\ \sigma_i^2 + 2W_0Q_i\ \text{cov}(R^*,R_i)\right] \tag{6}$$

where σ_*^2 and σ_i^2 are the variance of rate of return of the initial portfolio and of stock i. The following approximations which simplify the problem can be made:

$$U''(\bar{W}^*) = U''(\bar{W}). \tag{A.1}$$

$$\frac{U(\bar{W}) - U(\bar{W}^*)}{U'(\bar{W}^*)} \approx \bar{W} - \bar{W}^* \qquad (A.2)$$

The approximations are legitimate in terms of the process described in footnote 5.[6]

Using (A.1) and (A.2), (6) can now be written as;

$$\tfrac{1}{2}\frac{z}{W_0}\left[Q_i^2\ \sigma_i^2 + 2W_oQ_i\ \mathrm{cov}(R^*,R_i)\right] - \left[\bar{W} - \bar{W}^*\right] = 0 \qquad (7)$$

where $z = -\dfrac{U''(\bar{W}^*)W_0}{U'(\bar{W}^*)} \approx$ the Pratt index of relative risk aversion.

Note from (3) and (4) that

$$\bar{W} - \bar{W}^* = Q_i(\bar{R}_i - R_f) + C_i(1 + R_f) \qquad (8)$$

and, using the definition of $\tilde{R}^*$, that

$$\mathrm{cov}(R^*, R_i) = k\sigma_{ie} + \frac{Q_p\ \sigma_{ip}}{W_0} \qquad (9)$$

where σ_{ie} is the covariance between the rate of return on stock i and the rate of return on portfolio E, and σ_{ip} is the covariance between the return on stock i and the return on the initial trading account. Furthermore k, the desired holding of portfolio E (represented by point N in Figure 1) can be eliminated from (9) since it depends on the utility function and the known desired opportunity set (the line R_fE in Figure 1). By setting the slope of the dealer's indifference curve equal to the slope of the desired opportunity set, it can be shown that[7]

$$\bar{R} - R$$

Substituting (10) in (9) and (9) and (8) in (7) yields:

$$\frac{1}{2}\,\frac{z}{W_0}\,Q_i^2\,\sigma_i^2 + \frac{z}{W_0}\,Q_iQ_p\,\sigma_{ip} - C_i(1 + R_f) - Q_i\left[(\bar{R}_i - R_f) - (\bar{R}_e - R_f)\,\frac{\sigma_{ie}}{\sigma_e^2}\right] = 0 \qquad (11)$$

and

$$C_i = \frac{\frac{z}{W_o}\,\sigma_{ip}Q_pQ_i + \frac{1}{2}\,\frac{z}{W_o}\,\sigma_i^2Q_i^2 - Q_i\left[(\bar{R}_i - R_f) - (\bar{R}_e - R_f)\,\frac{\sigma_{ie}}{\sigma_e^2}\right]}{1 + R_f} \qquad (12)$$

Portfolio equilibrium for the dealer requires that:[8]

$$\bar{R}_i - R_f = (\bar{R}_e - R_f)\,\frac{\sigma_{ie}}{\sigma_e^2}\;. \qquad (13)$$

This result depends on Assumption 1--that the dealer can borrow and lend at R_f and assumes security i is in his investment account. Given (13), the last term in the numerator of (12) is zero. Note also that letting $R_f = 0$ in the denominator of (12) has minimal effect on (12).

These modifications yield the dollar holding cost function which is to be viewed as an incremental cost function since it refers to the cost of a single additional transaction undertaken in the trading interval:

$$C_i = \frac{z}{W_0}\sigma_{ip}Q_pQ_i + \frac{1}{2}\frac{z}{W_0}\sigma_i^2Q_i^2 \qquad (14)$$

The percentage cost is :

$$\frac{C_i}{Q_i} = c_i = \frac{z}{W_0}\sigma_{ip}Q_p + \frac{1}{2}\frac{z}{W_0}\sigma^2Q_i \qquad (15)$$

The holding cost of taking a position in stock i depends on :

(1) Dealer characteristics--relative risk aversion z, and dealer equity W_0. Of two dealers in the same stock, the one with larger z and/or smaller W_0 charges a higher fee for taking a position of given size ; or, at the same fee, would take smaller positions.

(2) Size of the transaction in stock i, Q_i. Total cost rises as the square of Q_i, and percentage cost rises linearly with Q_i.

(3) Characteristics of the stock--variance of return and the co-variance between the return on stock i and the return on the initial trading account portfolio[9]. Note that the covariance with the investment account does not enter.

(4) Size of the initial position in the trading account, Q_p. If Q_p is positive (and $\sigma_{ip} > 0$), the cost of buying stock i is larger than if there were no initial position. Conversely the cost of selling stock i is smaller than if there were no initial position.

In Figure 2, C_i is plotted as a function of Q_i using some reasonable values for the remaining variables. Placement of the curves depends on the dealer's initial position, Q_p, and the size of σ_{ip}. If $Q_p = 0$ or if $\sigma_{ip} = 0$, dollar cost has a minimum of $Q_i = 0$. If $Q_p \neq 0$ and $\sigma_{ip} > 0$, the minimum is at $Q_i \gtrless 0$ according as $Q_p \lessgtr 0$. A notable aspect of (14) illustrated by Figure 2 is its symmetry--a sale of given size costs the dealer the same amount as a purchase of given size. Assuming that the dealer has no initial holding in the trading account and is therefore at N in Figure 1, this result implies that points A and B in Figure 1, which represent long and short positions of the same amount, lie on the same indifference curve, as shown. Although B is much farther inside the efficient frontier than A, it is closer to the level of risk desired by the dealer and these two factors are offsetting. If the probability of purchases by the dealer equals the probability of sales by the dealer, the symmetric cost function implies that the optimal inventory in the dealer's trading account is zero; or, in other words, that the optimal overall portfolio of the dealer is the same as that of any nondealer with the same preferences and expectations. Thus even after becoming a dealer the desired portfolio remains point N in Figure 1.

Inability of the dealer to borrow and lend at the same rate of interest can eliminate the symmetry of the cost function and could complicate the problem slightly. In particular suppose the dealer can borrow at R_f but can lend the proceeds of short sales only at a fraction, θ, of R_f. This affects the third term in the numerator of (12) because θR_f replaces R_f. Given (13), the term does not go to zero but to $Q_i^s R_f(\theta - 1)$, where $Q_i^s < 0$ is the dollar value of short sales required and $\theta = 1$ if $Q_i^s = 0$ and

$\theta < 1$ if $Q_1^s < 0$.[10] The effect is to raise C_1 by the amount of interest not earned on the proceeds of a short sale. Dollar costs are therefore greater for $Q_1 < 0$ than for $Q_1 > 0$. Given a symmetric probability distribution on purchases and sales, the dealer tends, in this case, to keep a positive inventory in the trading account to avoid the extra cost of short selling[11].

It should be pointed out that θ can be made a more complicated function of other variables. For example if the bank is concerned about default risk on dealer borrowings, the logical variable is the dealer's debt-equity ratio, $L = \frac{Q_1 + Q_p}{W_0}$, where $\frac{d\theta(L)}{dL} > 0$. Such a modification is straightforward and would tend to raise dealer costs above those in (14). However, since the theoretical form and empirical justification for these modifications are not clear and since these modifications would not alter the character of the final model while adding to its complexity, the modifications are omitted from further consideration.

IV. Bid and Ask Prices

The dealer is compensated by purchasing shares at the bid price, p^b, usually below the "true" price, p^*, and by selling shares at the ask price, p^a, usually above the "true" price. Consider a dealer who specializes in a single stock i. Curve c_{i0} in Figure 3 represents the percentage holding cost function of the dealer with no initial position in the stock. Suppose he stands ready to buy $Q_{i0}^b > 0$ and sell $Q_{i0}^a < 0$, and quotes bid and ask prices that just cover his costs of doing so[12]. Then the price of immediacy of a sale of Q_{i0}^b to the dealer is set at $\frac{p_{i0}^* - p_{i0}^b}{p_{i0}^*} = c_{i0}(Q_{i0}^b)$, and the price of immediacy of buying from the dealer is $\frac{P_{i0} - P^a_{i0}}{P_{i0}} = c_{i0}(Q^a_{i0})$. If a seller appears and trades Q^b_{i0} at P^b_{i0}, the dealer sets the new ask price in the second (and final) trading interval at the new equilibrium price, $P^a_{i,t+1} = P^e_{i,t+1}$, and the position is sold. If a buyer appears and trades Q^a_{i0} at P^a_{i0}, the dealer sets $P^b_{i,t+1} = P^e_{i,t+1}$ in the second trading interval, and the short position is covered.

A dealer faced with a trade of Q_i would like in the same trading interval to make an offsetting transaction, in the same stock or some equivalent combination of stocks, that perfectly hedges his portfolio. Real world constraints imposed in this model are that the dealer cannot actively and immediately take such offsetting positions in the same stock or other stocks. However, the bid-ask quotation in any stock is set so as to encourage transactions which reduce the risk of holding the initial portfolio. This point is illustrated in Figure 3 by the line c_{i1}, the percentage cost function for $Q_p > 0$, $\sigma_{ip} > 0$. Since in this case the dealer already has a long position the return on which is positively correlated with stock i's return, the bid and ask prices are set so as to encourage sales by the dealer of stock i and to discourage purchases by the dealer of stock i. Thus the bid price, P^b_{i1}, is lower than if there were no initial position and the ask price, P^a_{i1}, is higher. If $Q_p = Q_i$, $c_i = -\frac{1}{2}\frac{z}{W_0}\sigma^2_i Q_i$; and in this case the dealer is willing to pay customers to take him back to his optimum portfolio (N in Figure 1) an amount equivalent to the cost of holding the risky position. If it is assumed he has been paid for the immediacy costs of trading Q_p, the dealer just breaks even and is properly compensated during each time interval for bearing risk.

The percentage spread is the percentage difference between bid and ask price, or just the vertical distance between two percentage prices in Figure 3. The spread function corresponding to (15) for $Q_i^b = Q_i^a = |Q_i|$ is :

$$s_i = \frac{P_i^a - P_i^b}{P_i^*} = c_i(Q_i^b) - c_i(Q_i^a) = \frac{z}{W_0}\sigma_i^2 \, |Q_i| \qquad (16)$$

which is independent of the initial inventory of the dealer and does not involve any covariance term. Thus, if the dealer prices just to cover the costs of each transaction, the spread (but not the bid or ask price) is independent of the initial inventory and therefore holds for the dealer in many stocks as well as for a dealer in one stock. This result depends on the assumptions that previous inventory holding costs are sunk costs and that only one stock is traded per trading interval.

V. Other Costs

In addition to holding costs the dealer incurs other costs that shall receive relatively brief treatment here. First the dealer incurs certain explicit costs--called order costs--in carrying out a transaction. These costs which are incurred by brokers as well as by dealers include the cost of labor, the cost of communicating, and the cost of clearing and record keeping. The simplest assumption is that order costs are a constant dollar amount, M, per transaction and therefore a declining proportional amount $\frac{M}{\ }$ per dollar

A second cost arises if some investors trade with the dealer because they have superior information. [See Bagehot (1971). Jaffe & Winkler (1976) on this point.] In organized markets, a dealer quotes a bid price at which he is willing to buy and an ask price at which he is willing to sell without knowing whether the next trade will be a purchase or sale. Even if the dealer possesses inside information (because of knowledge of the book of limit orders, for example), such information would be reflected in bid and ask prices; and relative to such knowledge he is still subject to losses from investors who have information he does not possess. On the assumption that the dealer cannot distinguish information traders from others without information (liquidity traders), the dealer must increase his bid-ask spread vis-à-vis all traders to protect himself against possible losses of dealing with information traders. He widens the spread so as to recover from liquidity traders what he loses to information traders.

For the purposes of this paper, information trading can be incorporated into the dealer's cost function by recognizing that (13) does not hold in such a case. Instead :

$$\bar{R}_i - R_f = (\bar{R}_e - R_f)\,\frac{\sigma_{ie}}{\sigma_e^2} - \gamma a_i \qquad (17)$$

where a_i is the expected return on the information possessed by those that trade with the dealer, assumed the same for buyers and sellers and independent of transaction size, and

$$\gamma = \begin{cases} 1 & \text{if dealer purchases shares} \\ -1 & \text{if dealer sells shares .} \end{cases}$$

In other words the dealer expects to earn less than he would in the absence of informational trading, and his bid-ask spread must be wider as a result.[13]

In addition informational trading may cause σ_i^2 and σ_{ip} to be different, but this is not assumed here. The effect of (17) is to make the last term in the numerator of (12) nonzero and equal to $-\gamma a_i$.

Modifying the percentage cost function (15) to include order costs and information costs yields:

$$c_i = \frac{z}{W_0}\sigma_{ip}Q_p + \frac{1}{2}\,\frac{z}{W_0}\sigma_i^2 Q_i + \gamma a_i + \frac{M}{Q_i} \tag{18}$$

The adverse information cost (a_i) and order cost $\frac{M}{Q_i}$ could undoubtedly be treated in a more complex way. For example the likelihood of adverse information may be a function of the size of the transaction (because the rich are smarter). This would result in higher bid-ask spreads as Q_i increases not only because of larger holding costs but also because of larger information costs. In principle the two costs could be distinguished because information costs are not a function of dealer inventory level whereas holding costs are. Similarly order costs may in part be a function of transaction size. This would result in a less steeply declining per dollar order cost as a function of transaction size.

Such modifications are unlikely to change the basic shape of the cost function (18) which is plotted on Figure 4. The difference between Figures 4 and 3 is that there is now a discontinity at $Q_i = 0$ due to switches in sign of γ and the presence of M.

A second more important difference is that there is now an optimal scale for the dealer because falling order costs are offset by rising holding costs. Adverse information costs do not affect the scale decision because they are assumed independent of transaction size. If the dealer operates at minimum average cost, [the minimum of (18)] output is at:

$$Q_i^* = \pm\sqrt{\frac{2\ MW_0}{z\ \sigma_i^2}}\ , \tag{19}$$

also shown in Figure 4. Whether the dealer actually operates here depends on the number of other firms making a market in the stock and on the existence of special skills of the dealer, etc.

A spread function independent of Q_p corresponds to (18) :

$$s_i = \frac{z\sigma_i^2}{W_0}\ |Q_i| + 2a_i + \frac{2\cdot M}{|Q_i|} \tag{20}$$

The spread function is a "U" shaped function of $|Q_i|$.

VI. Organization of Dealers

An important policy issue is the organization of dealers that regulatory policy should foster--whether a monopolistic specialist system such as has existed on the New-York Stock Exchange where no stock has had more than one dealer or a competitive dealer system such as has been permitted in Over-the-Counter stocks.[14]

Probably the principal undesirable aspect of monopoly dealers is pricing above marginal cost, the extent of which depends on the elasticity of demand for immediacy, a subject not considered here. However the organization of dealers also affects the cost of dealer services.

First consider the effects of changing from monopoly dealers to competitive dealers in the short run where the number of dealers is fixed. In a free market a given number of dealers allocate themselves among stocks such that the cost of the marginal transaction depends only on characteristics of stocks and not on characteristics of dealers. Wealthier dealers take larger positions (in the form of holding more stocks or more per stock) than less wealthy dealers so that marginal cost in any particular stock is equal across dealers. Similarly dealers less averse to risk take larger positions than more risk averse dealers. To the extent that the allocation of stocks to dealers under a monopoly dealer system is not optimal, reorganization into a competitive system produces a net welfare gain to society by equalizing marginal costs across dealers. Reorganization into a competitive system benefits markets in stocks with previously inadequate capital or high z specialists and harms markets in stocks with previously excessive capital or low z specialists.[15]

Second, if entry restrictions cause monopoly dealers to operate at too large a scale (to the right of Q_i^* in Figure 4), competition from new entrants would push costs to the minimum. The equilibrium number of dealers in a stock can be derived using (19) if one is willing to assume that dealers have identical cost functions and operate at the minimum average cost. Let $|D_i|$ denote the absolute dollar value of investors' trading with dealers in stock i at the price of immediacy corresponding to minimum average cost.

Let $|Q_i^*|$ = absolute dollar value of minimum cost output. The equilibrium number of dealers, d_i^*, is the number that demand can support if all are operating at minimum average cost. From (19) this is:

$$d_i^* = \frac{|D_i|}{|Q_i^*|} = |D_i| \cdot \left| \sqrt{\frac{z\ \sigma_i^2}{2\ M\ W_0}} \right| . \tag{21}$$

For given long-run demand, the number of dealers is greater in riskier stocks and stocks in which individual dealers are more risk averse; the number is less the greater the order costs incurred by dealers and the greater the wealth of the individual dealer.

It is interesting to note a few policy implications of (21). It is sometimes suggested that under competition, dealers in risky stocks would not be forthcoming. However, (21) suggests that, ceteris paribus, more dealers will operate in risky stocks. This is due to the fact that each dealer will take a smaller position. Another policy question concerns dealer capital requirements (our W_0). Regulators argue that dealers should be required to maintain a minimum capital in each stock.[16] Under monopoly dealerships such requirements may be useful in ensuring adequate capital since entry of dealers with additional capital is prohibited. (The appropriate level of capital depends on the characteristics of the stock and the dealer and would be quite difficult to set precisely.) Under competitive markets, minimum capital requirements may be counterproductive; if set too high, they reduce the number of dealers in a stock according to (21) and thereby reduce the beneficial effects of competition. (Furthermore, minimum capital requirements may have little effect in practice because it is difficult to compel utilization of capital.)

A final issue concerns the ease of entry and exit under a competitive system. To the extent that entry and exit costs are zero, dealers would fluctuate so that supply is perfectly elastic at the price of liquidity corresponding to $|Q_i^*|$ even in the short run. If short-run supply is derived for a period in which the number of dealers in each stock is fixed, it is upward sloping. This is realistic because they are probably entry costs and because regulations often require dealers to maintain a market for a minimum period of time (6 months in the O.T.C.). Such regulations are undesirable in that they lengthen the short-run and raise costs on average.

VII. Multiperiod Considerations

There is no guarantee that the dealer can readily liquidate his inventory in the second trading interval. Assume there is a cost, $\tilde{D}_i$, of liquidating the inventory should it still exist. This cost, a random variable in the first trading interval, can be thought of as the payment necessary to cause someone to accept the inventory at the new "true" price, or alternatively as the difference between trade price (bid or ask price) and the new "true" price that is necessary to create sufficient incentive to others to purchase the dealer's inventory. Under certain assumptions one can view $\tilde{D}_i$ as the (implicit) payment by the dealer to himself that makes him willing to continue to hold the inventory.

It is helpful to think of events unwinding in the following time sequence. In the first trading interval the dealer takes a position, Q_i. He enters a period of price uncertainty in which there is no trading. The second trading interval is divided into two parts. The dealer enters the second trading interval by pricing his inventory at P^*_{t+1}, the new "true" price. Thus if he initially bought the stock (at P^b_{it}), he sets $P^a_{i,t+1} = P^*_{i,t+1}$; if he initially sold the stock (at $P^a_{i,t}$), he sets $P^b_{i,t+1} = P^*_{i,t+1}$. In the model of Section II, the world ends here, and the dealer is assumed to liquidate his position at $P^*_{i,t+1}$. In fact he may not be able to liquidate his inventory at $P^*_{i,t+1}$. If it is not liquidated, he sets at new concession price which deviates from $P^*_{i,t+1}$, and which is sufficient to liquidate his position. The dollar amount of the concession on all his shares is $\tilde{D}_i$, a random variable in the first trading interval.

The one period framework of section II can be maintained but modified simply by noting that terminal wealth in (4) will be reduced by $\tilde{D}_i$. The development proceeds exactly as before except that some of the expressions are more complicated. The primary complication arises in going from (5) to (6), which involves writing $E\,(W-\bar{W})^2 = \sigma^2(W)$ in terms of its components. With $\tilde{D}_i$ this is

$$\sigma^2(W) = W_0^2\,\sigma_*^2 + Q_i\,\sigma_i^2 + \sigma^2(D_i) + 2W_0Q_i\,\mathrm{cov}(R^*,R_i) - 2W_0\,\mathrm{cov}(R^*,D_i) - 2Q_i\,\mathrm{cov}(R_i,D_i) \qquad (22)$$

Recall that $\tilde{D}_i$ is the concession relative to the "true" price and that R and R^* are returns based on "true" prices. There is no reason to believe that $\tilde{D}$ should be correlated with these returns. Thus :

$$\mathrm{cov}(R^*, D_i) = \mathrm{cov}(R_i, D_i) = 0 \tag{A.3}$$

With these assumptions and otherwise making the same assumption as in section II, the simplified cost function becomes:

$$C_i = \frac{z}{W_0}\sigma_{ip}Q_pQ_i + \bar{D}_i + \frac{1}{2}\frac{z}{W_0}\sigma^2(D_i) + \frac{1}{2}\frac{z}{W_0}\sigma_i^2Q_i^2 \tag{23}$$

Consider now a more precise specification of $\tilde{D}_i$ that is based on the assumption that there is never partial liquidation of any prior transaction. This implies that $\tilde{D}_i$ is a Bernoulli variate taking the value zero if Q_i is sold at $P^*_{i,t+1}$ and the value $\hat{C}_i$ if the position is not sold. The corresponding probabilities are $(1-\pi_i)$ and π_i. Then :

$$\bar{D} = \pi_i\hat{C}_i \tag{24 a}$$

$$\sigma^2(D) = \pi_i(1-\pi_i)\hat{C}_i \tag{24 b}$$

Now suppose $\hat{C}_i = C_i$, that the price concession necessary to eliminate a position equals the cost of assuming the position in the first place. Substituting C_i for $\hat{C}_i$ in (24) and substituting for $\bar{D}$ and $\sigma^2(D)$ in (23) yields

$$C_i = \frac{\frac{z}{W_0}\sigma_{ip}Q_pQ_i + \frac{1}{2}\frac{z}{W_0}\sigma_i^2Q_i^2}{(1-\pi_i)\left[1 - \frac{1}{2}\frac{z}{W_0}\pi_i\right]} \tag{25}$$

If the probability of a forced liquidation is zero, the cost function is

The solution (25) can be stated in multiperiod terminology by letting C_i represent the cost to the dealer of continuing to hold the position acquired in the first period. Under this interpretation, the dealer does not liquidate his inventory at a concession price, but he continues to price it at the "true" price and to wait until an investor trades in the opposite direction. This is not a true multiperiod framework in which intermediate decisions would be allowed. It is the case that $\hat{C}_i = C_i$ under the assumption that characteristics of the stock (σ_i^2, σ_{ip}) and the dealer (z, W_0) are unchanged overtime and that the probability (π) of holding the stock is unchanged. Independance and stationarity of the distribution of returns and of trading volume will lead to unchanged $\sigma_i, \sigma_{ip}, \pi_i$ for given Q_p. It is not unreasonable to assume constant z. However W_0 changes, and we must argue that the change is too small to have a significant effect.[17]

When viewed as the cost of continuing to hold the stock, it is natural to specify the dealer's cost in terms of the number of periods he expects to hold the inventory when it is priced at $\overset{*}{P}$. Under the assumption of the stationarity and independence of the distribution of volume, there is a simple relationship between the expected holding period τ_i and π_i :

$$\tau_i = 1 + \sum_{h=1}^{T} \pi_i^h \, h(1-\pi_i) \qquad (26)$$

where h = number of periods inventory is held

T = total number of possible periods.

Letting T → ∞ (26) can be shown to be :

$$\tau_i = \frac{1}{1-\pi_i} \qquad (27)$$

Then, from (25)

$$C_i = \frac{\frac{z}{W_0} \tau_i \sigma_{ip} Q_p Q_i + \frac{1}{2} \tau_i \frac{z}{W_0} \sigma_i^2 Q_i^2}{1 - q_i} \qquad (28)$$

where

$$q_i = \frac{1}{2} \frac{z}{W_0} \left(\frac{\tau_i - 1}{\tau_i} \right)$$

This differs from (14), only in that the expected holding period multiplies the per-period variance and covariance and in the second term in the denominator, which is small. If $\tau_i = 1$, (14) results.

VII. Summary and Conclusions

A dealer cost function composed of holding costs, order costs and information costs is developed. The emphasis is on the holding cost component which is derived on the assumption that the cost is an amount which maintains the dealer's level of expected utility of terminal wealth in response to transactions imposed upon him by the public that tend to move him away from his optimal portfolio. The holding cost depends on the dollar size of the transaction, the variance of return of the stock being traded, the size of the initial holdings of all stocks in the dealer's trading account, the covariance between the return on the stock being traded and the return on the trading account, the wealth of the dealer, and his attitude toward risk. Dollar holding cost for the incremental trade is a quadratic function of the size of the position acquired, and percentage holding cost in thus linear. Under the assumption that dealers are able to earn full interest on the proceeds of short sales, the cost function is symetric--that is, the cost of going short equals the cost of going long. Under certain simplifying assumptions, the multiperiod holding cost function is shown to be quite similar to the one period function, differing only in that the holding period enters the function (or, equivalently in this model, the probability that the dealer is unable to dispose of his inventory at the equilibrium price after one period).

The order cost is a minimum cost per transaction which therefore declines per dollar as the size of the transaction increases. Falling order costs and rising holding costs determine an optimum scale of operation by the dealer in each of his stocks. Information costs arise when investors trade on the basis of superior information, which adversely affects the dealer's expected return on his inventory.

The paper also examines some policy issues related to the organization of dealers. Monopoly dealers are undesirable not only for the standard reason that they price above marginal cost but also because the assigment of stocks to dealers may be arbitrary and can result in a nonoptimum distribution of dealer wealths and risk attitudes across stocks and because limits on entry may cause dealers to operate at a nonoptimal scale. The equilibrium number of dealers in a comparative system is also considered. Given demand, the number of dealers increases with the risk of the stock and the risk aversion of the "representative dealer". The desirability of regulating minimum capital requirements and imposing other entry conditions is questioned.

FOOTNOTES

[1] For the purposes of this paper investors include individuals and institutional investors who in other contexts might be considered to be intermediaries.

[2] This assumption is realistic for dealers that also do a brokerage business (as in the OTC market) and have on deposit customer securities, that can be borrowed at no cost. It is also realistic if aggregate borrowings of stock are small since competition by lenders (owners) of stock would tend to drive down the cost to the borrower of stock. Discussions with nonbroker dealers suggest that there is at present a sharing between borrower and lender of the interest on the proceeds of short sales. Recognition of this fact changes somewhat the results of the model and the implications are discussed at the appropriate point in the paper.

[3] A possible explanation for the observed tendency to find noncorporate forms of organization in the securities industry is the frequency of verbal committments, the fulfillment of which depends on personal integrity and the threat of personal bankruptcy.

[4] The dollar value of borrowing or lending differs from the "true" value of the transaction exactly by the cost, C_i, only in perfect competition. When competition is less than perfect, the dealer can price above cost, and therefore the discount from "true" value exceeds cost.

[5] Dropping these higher order terms can be justified by assuming that the price dynamics for stocks is given by

$$R_i = \mu_i \Delta t + \sigma_i \sqrt{\Delta t} \; \tilde{Z}$$

Where R_i = rate of return during the interval Δt.
μ_i = expected rate of return during Δt.
σ_i = standard deviation of return during Δt.
$\tilde{Z}$ = normal random variable with zero mean and unit variance.
Δt = time interval.

Dropping terms of order higher than two implies that terms in Δt raised to powers of 2 or more are dropped, which is justified since Δt is assumed to be very small in this analysis : let $\tilde{W} = W_0(1+\tilde{R})$. In terms of the above process

$$\tilde{W} = W_0(1 + \mu\Delta t + \sigma\sqrt{\Delta t}\; \tilde{Z})$$

Then in (5)

$$(\tilde{W}-\bar{W}) = W_0\sigma \sqrt{\Delta t} \; \tilde{Z}$$

$$(\tilde{W}-\bar{W}^*) = W_0\sigma_*\sqrt{\Delta t} \; \tilde{Z}$$

Therefore terms of order three would involve $(\Delta t)^2$.

[6] On the basis of the process in the preceding footnote, (5) can be written as

$$U(\bar{W}^*) + \frac{1}{2} U''(\bar{W}^*) \; W_0^2 \; \sigma_*^2 \, \Delta t = U(\bar{W}) + \frac{1}{2} U''(\bar{W}) \; W_0^2 \; \sigma^2 \; \Delta t \qquad (F6.1)$$

Now consider A.1 and expand $U''(\bar{W}^*)$ around $\bar{W}$ by one term :

$$U''(\bar{W}^*) = U''(\bar{W}) + (\bar{W}^* - \bar{W}) \; U'''(\bar{W}). \qquad (F6.2)$$

Substituting for $(\bar{W}^* - \bar{W})$ on the basis of the process in fn. 5 and substituting (F7.2) in (F7.1) will show that neglecting the second term in (F7.2) neglects a term in $(\Delta t)^2$ which is small. Following a similiar procedure for A.2, one gets

$$U(\bar{W}) - U(\bar{W}^*) = (\bar{W} - \bar{W}^*)U'(\bar{W}^*) + \frac{1}{2}(\bar{W} - \bar{W}^*)^2 U''(\bar{W}^*) \qquad (F6.3)$$

Or

$$U(\bar{W}) - U(\bar{W}^*) = (\bar{W}-\bar{W}^*)\,U'(\bar{W}^*) + \frac{1}{2}W_0(\mu-\mu_*)^2(\Delta t)^2\,U''(\bar{W}^*) \qquad (F6.4)$$

The last term in (F7.4) involves $(\Delta t)^2$ and can thus be neglected.

[7] For $Q_p = 0$, terminal wealth is:

$$\tilde{W}^* = W_0[1 + R^*] = W_0[1 + k\tilde{R}_e + (1 - k)R_f] .$$

Using this definition of $\tilde{W}^*$, the differential of the L. H. S. of (6) is:

$$dEU(W^*) = \frac{\partial U(W^*)}{\partial \bar{W}^*} \cdot \frac{\partial \bar{W}^*}{\partial \bar{R}^*}\, d\bar{R}^* + U''(\bar{W}^*)W_0^2\sigma_* d\sigma_* .$$

Note that $\frac{\partial \bar{W}^*}{\partial \bar{R}^*} = W_0$ and $\sigma_* = k\sigma_e$, set the differential equal to zero and solve for the slope of the indifference curve:

$$\frac{d\bar{R}^*}{d\sigma_*} = -\frac{U''(\bar{W}^*)}{U'(\bar{W}^*)}W_0 k\sigma_e .$$

Setting this equal to the slope of the opportunity set, $\frac{\bar{R}_e - R_f}{\sigma_e}$, yields (10).

[8] Under homogeneous expectations this would be the Sharpe security market line, an equilibrium relationship, with portfolio E being the market portfolio and σ_{ie}/σ_e^2 the well-known "Beta" coefficient.

[9] Note that σ_i^2 and σ_{ip} are not directly observable since they depend on variability in "true" returns. Observed variability of return depends as well on the cost of immediacy which is reflected in bid and ask prices and which in turn depends on the volume of trading and other variables.

[10] Since the dealer may have stock i in his trading account, short sales may not be necessary when the dealer sells stock i.

[11] The observed tendency to find positive inventory may be due to a number of other factors. For example dealers may be able to anticipate buying by the public better than selling by the public. Second, and probably most important, there are often tax benefits to carrying stock in one's trading account rather than one's investment account. If the dealer is taxed as a corporation, the corporate tax may be lower than his individual tax. Long term losses can receive ordinary income tax treatment in the trading account.

[12] It is assumed that the dealer is in a competitive environment for the purposes of this illustration.

[13] a_i is analogous to the Jensen (1968) measure of mutual fund performance.

[14] As of the Spring 1976 competing dealers were permitted on the NYSE, but few have as yet appeared.

[15] However, the total gain of shifting from a monopoly to a competitive system may not be as great as implied here. For example, monopoly profits may make it possible for regulatory authorities to require dealers to act as if they had more capital or a smaller z.

[16] On the NYSE, the specialist must be able to carry 2,000 shares of each stock (e.g., \$ 80,000 for shares priced at \$ 40). However, most of the funds for carrying inventory can be borrowed. On the OTC a dealer must have net capital of \$ 2,000.

[17] On average W_0 increases over time, which would reduce costs. This partly offsets the preponderence of factors which lead to increased costs.

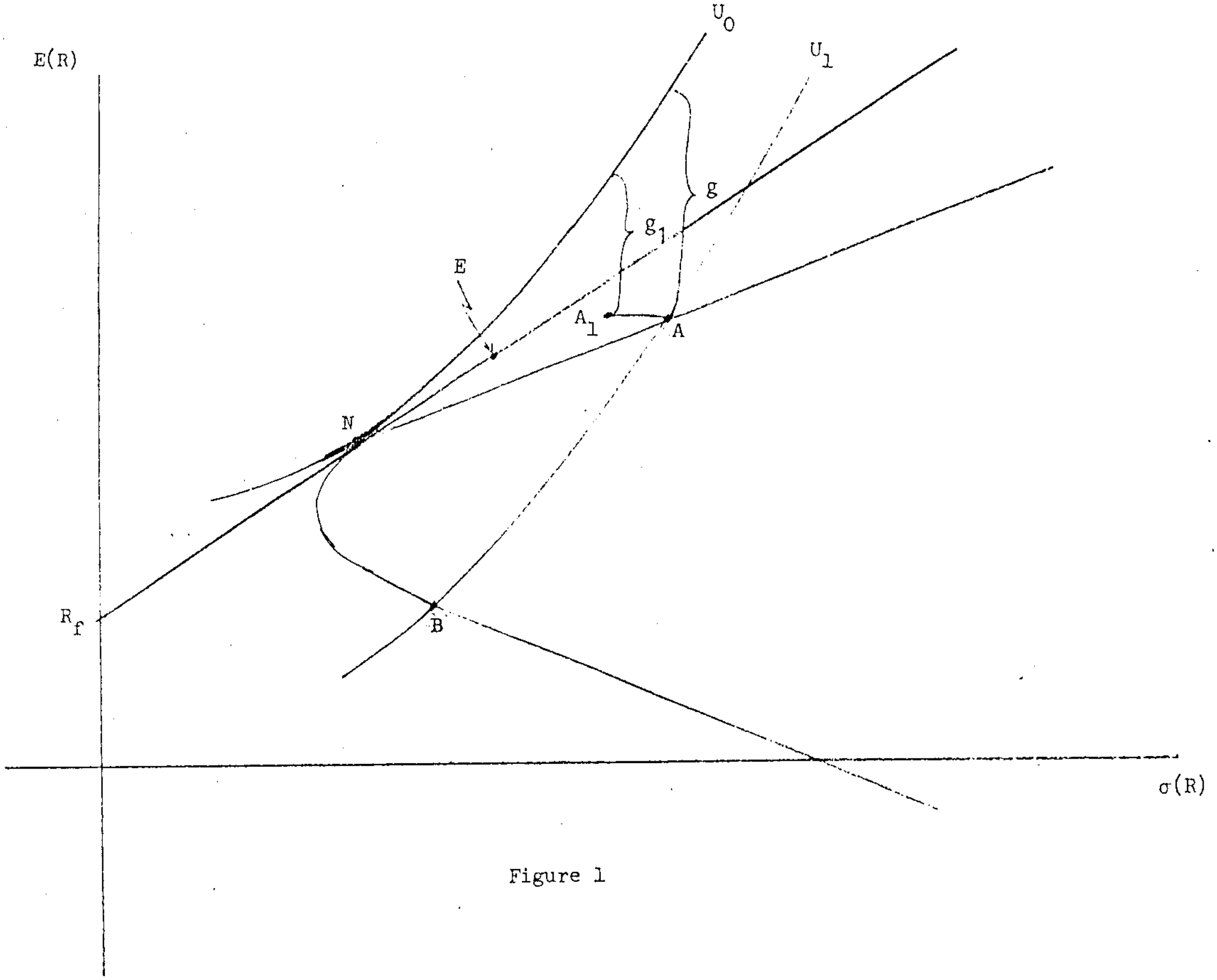


Figure 1

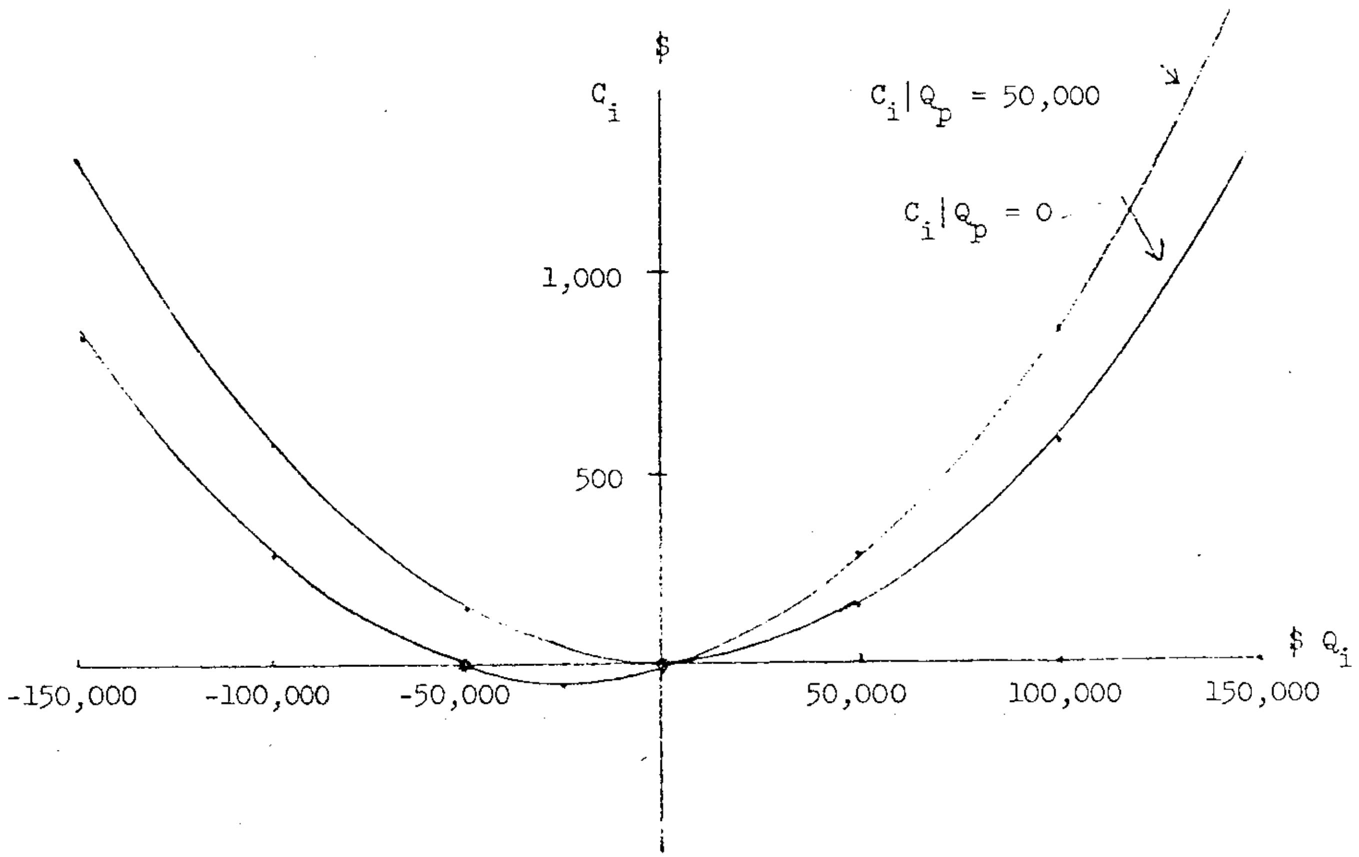
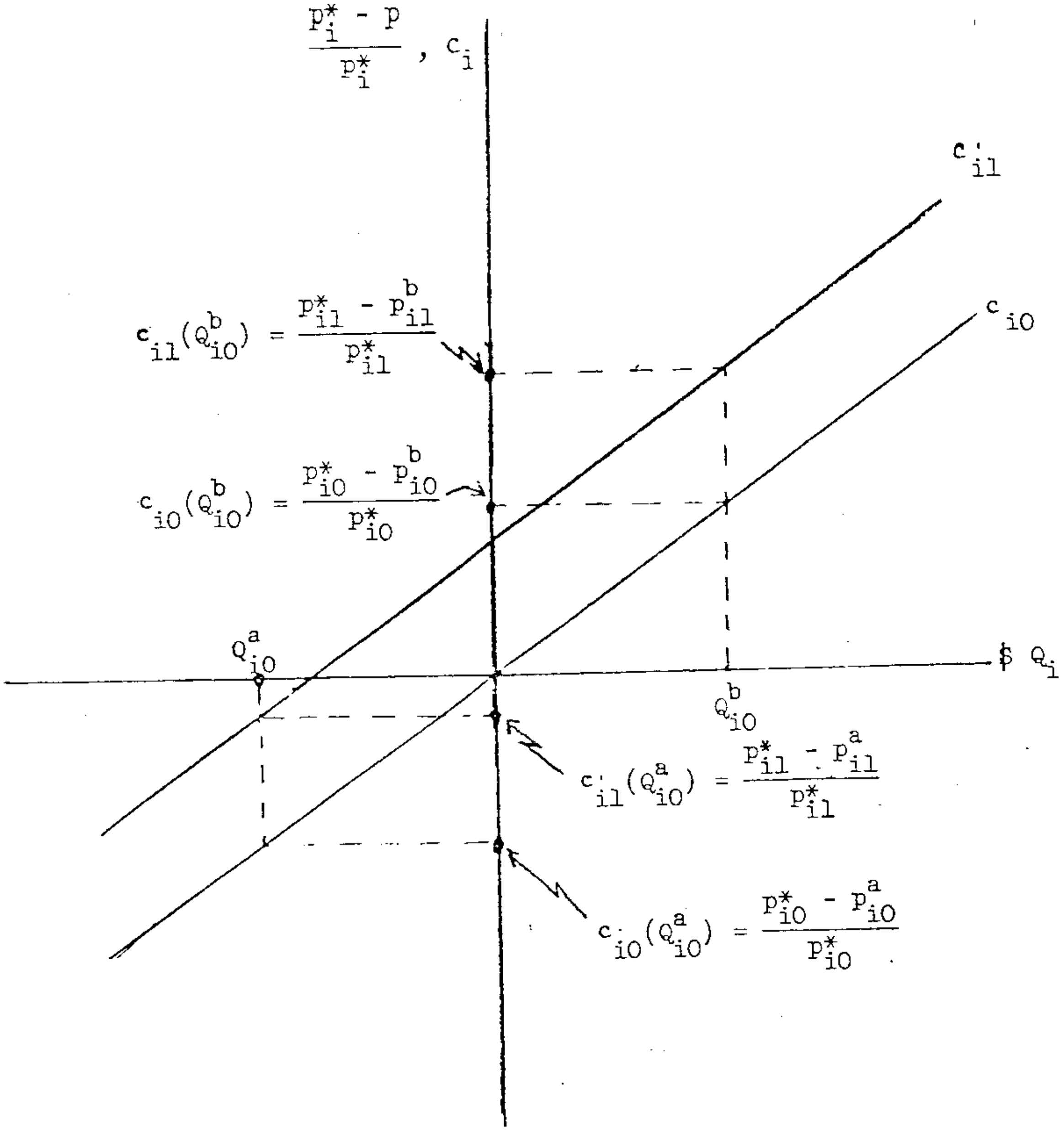

$
C_i
$C_i | Q_p$ = 50,000
$C_i | Q_p$ = 0
1,000
500
$ Q_i
-150,000
-100,000
-50,000
50,000
100,000
150,000

Figure 2



$\frac{p_i^* - p}{p_i^*}$, c_i
c_{i1}
c_{i0}
$c_{i1}(Q_{i0}^b) = \frac{p_{i1}^* - p_{i1}^b}{p_{i1}^*}$
$c_{i0}(Q_{i0}^b) = \frac{p_{i0}^* - p_{i0}^b}{p_{i0}^*}$
Q_{i0}^a
$ Q_i$
Q_{i0}^b
$c_{i1}(Q_{i0}^a) = \frac{p_{i1}^* - p_{i1}^a}{p_{i1}^*}$
$c_{i0}(Q_{i0}^a) = \frac{p_{i0}^* - p_{i0}^a}{p_{i0}^*}$

Figure 3

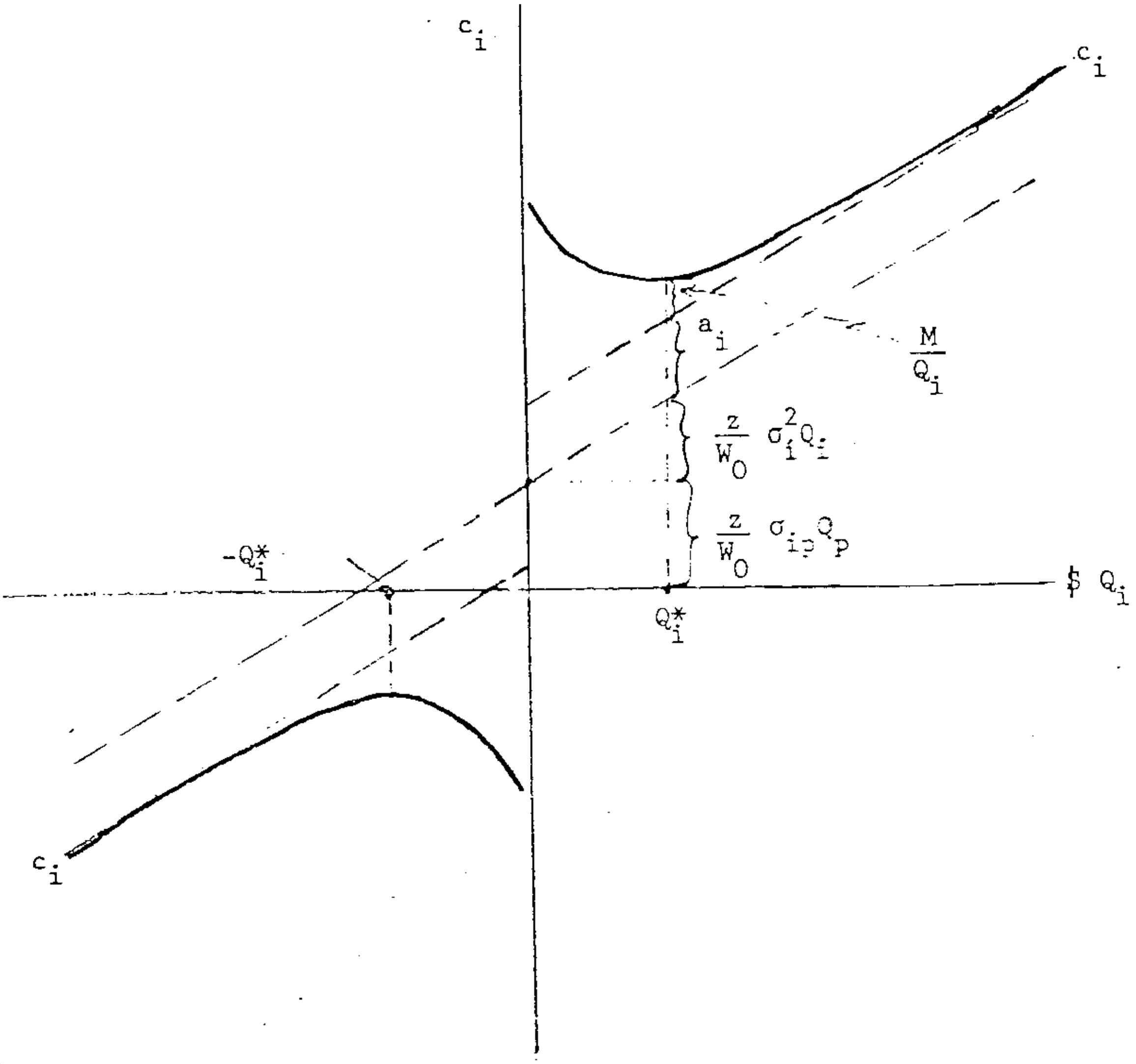

c_i
c_i
a_i
$\frac{M}{Q_i}$
$\frac{z}{W_0}\ \sigma_i^2 Q_i$
$\frac{z}{W_0}\ \sigma_{ip} Q_p$
$-Q_i^*$
$\$\ Q_i$
Q_i^*
c_i

Figure 4

REFERENCES

1. Bagehot, W., (pseud.), "The Only Game in Town," Financial Analysts Journal (March/April 1971).

2. Benston, George J. and Hagerman, Robert L., "Determinants of Bid-Asked Spreads in the Over-the-Counter Market," Journal of Financial Economics, 1 (Dec., 1974), pp. 353-364.

3. Copeland, Thomas E., "A Model of Asset Trading Under the Assumption of Sequential Information Arrival," Journal of Finance (September, 1976).

4. Demsetz, Harold, "The Cost of Transacting," Quarterly Journal of Economics, 82, No. 1 (February 1968).

5. Epps, Thomas W., "The Demand for Brokers' Services: The Relation Between Security Trading Volume and Transaction Cost," Bell Journal of Economics, 7 (Spring 1976).

6. Institutional Investor Study Report of the Securities and Exchange Commission, 92nd Congress, 1st Session House Document No. 92-64, March 12, 1971. Washington: G.P.O., 1971. Part 4, Ch. 12.

7. Jaffe, J. and Winkler, R., "Optimal Speculation Against an Efficient Market," Journal of Finance, 31 (March 1976).

8. Jensen, Michael C., "The Performance of Mutual Funds in the Period 1945-1964," Journal of Finance, 23 (May 1968).

9. SEC, "Policy Statement of the Securities and Exchange Commission on the Structure of a Central Market System," in Securities Regulations and Law Report, No. 196, April 4, 1973, published by Bureau of National Affairs, Inc.

10. Stoll, H. R., "The Pricing of Security Dealer Services: An Empirical Study of NASDAQ Stocks," forthcoming in Journal of Finance.

11. Tinic, Seha M., "The Economics of Liquidity Services," Quarterly Journal of Economics, 86 (February 1972).

12. Tinic, Seha M., and West, Richard, "Competition and the Pricing of Dealer Services in the Over-the-Counter Market," Journal of Financial and Quantitative Analysis, 7 (June 1972).