Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread

Kalman J. Cohen and Steven F. Maier

Duke University

Robert A. Schwartz

New York University

David K. Whitcomb

Rutgers University

By considering investor order placement strategy, this paper demonstrates that transaction costs cause bid-ask spreads to be an equilibrium property of asset markets. With transaction costs, the probability of a limit order executing does not go to unity as the order is placed infinitesimally close to a counterpart market quote; thus, with certainty of execution at the counterpart market quote, a "gravitational pull" is generated that keeps counterpart quotes from being placed infinitesimally close to each other. An equilibrium spread is defined and its size linked to market thinness; implications are noted for the design of a trading system.

I. Introduction

This paper establishes that transaction costs in secondary asset markets cause individual investors to use order placement strategies that

We thank Amir Barnea, Avraham Beja, Richard Burton, Mark Eaker, Dan Galai, Kenneth Garbade, Thomas Ho, Wesley Magat, David Peterson, and Marshall Sarnat for their helpful comments. Earlier drafts of this paper have been presented at seminars in economics and finance at Duke University, the Hebrew University at Jerusalem, New York University, Tel-Aviv University, and the Joint National Meetings of TIMS/ORSA (New York City, May 1978).

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

result in a nontrivial market bid-ask spread.[1] We define an *equilibrium* market spread and demonstrate that it will be greater for thinner securities.

The analysis fits into a growing body of literature which increasingly is being referred to as the microstructure of security markets. Stigler (1964), Demsetz (1968), West and Tinic (1971), Tinic (1972), and Tinic and West (1972, 1974), along with Farrar, Smidt, Stoll, and others involved in the institutional investor study (see U.S. Securities and Exchange Commission 1971), were among the first to focus rigorous analytical attention on the operations of security markets. More recently, microstructure theory has been explicitly considered by, among others, Garman, who coined the term (1976), Beja and Hakansson (1977), and Cohen, Maier, Schwartz, and Whitcomb (hereinafter CMSW) (1978). The analytical issues addressed involve the interplay among market participants, trading mechanisms, and the dynamic behavior of security prices. In the present paper, we study the formulation of optimal investor trading strategies and how these interact with one aspect of the dynamic behavior of security prices, the bid-ask spread. Spreads are of concern to investors because they are a variable cost of trading by market order and because they cause an inflation of transaction-to-transaction returns variance.

The pioneering analysis of bid-ask spreads was provided by Demsetz (1968). Further studies include: Tinic (1972), Tinic and West (1972, 1974), Benston and Hagerman (1974), Garman (1976), Hamilton (1976, 1978), Branch and Freed (1977), Stoll (1978*a*, 1978*b*), Ho and Stoll (1979, 1980), Newton and Quandt (1979), Schleef and Mildenstein (1979), Smidt (1979), and Amihud and Mendelson (1980). In CMSW (1979) we discuss this literature and contrast it with the formulation presented here. Except for Ho and Stoll (1980), there has been no explicit consideration of the transition from individual spreads to the market spread. It will be clear from our analysis that this transition is not a simple aggregation process and that the market spread is the product of a dynamic interaction involving many market participants. Also, previous theoretical models have generally assumed that investors can be dichotomized into two groups—immediacy demanders and immediacy suppliers. Our model of investor order placement strategy suggests that such a dichotomy will not be observed in the marketplace.

For the market, the spread is the difference between the lowest ask and the highest bid of all participants. In markets composed of many

[1] We have elsewhere (Cohen et al. 1980) considered how the impact of transaction costs on stock price movements introduces serial correlation in returns data and causes estimates of the market model beta coefficient to be biased.

traders with heterogeneous beliefs and trading propensities, one might expect to have orders at virtually every permissible price in the neighborhood of equilibrium and hence to find no significant market spread.[2] However, we show that even when expectations and trading propensities are heterogeneous, the spread is a property of asset markets that have temporarily cleared.[3] The analysis yields an existence proof of a noninfinitesimal spread with continuous pricing; a fortiori this proves for discrete prices the existence of a positive spread for nontrivial reasons.[4]

The essence of our argument is as follows. At any point in time, any investor might alternatively seek to trade via a limit order (be an immediacy supplier), trade with certainty via a market order (be an immediacy demander), or not seek to trade at all. Limit orders create the book, and market orders clear out limit orders. Because execution via market order is certain, while execution via limit order is not, it never pays for any investor to place a limit order (e.g., a bid) at a price too close to that of a counterpart limit order (e.g., an ask). Intuitively stated, as a trader contemplates placing a bid closer and closer to an ask already established on the market, he is increasingly attracted by this counterpart offer; at some point, the "gravitational pull" exerted by the established ask will dominate. The trader will "jump" his price and execute with certainty via a market order.

Section II establishes the scenario for our analysis. Section III focuses on the probability of a limit order executing and shows that, with transaction costs, this probability does not go to unity as the price at which a limit order is placed becomes infinitesimally close to a counterpart market quote. This demonstration underscores an investor's need for an order placement strategy and provides the foundation for our gravitational pull model. Section IV models the investor's order placement decision and develops conditions under which he will transmit limit or, alternatively, market orders to the market, or do nothing. The analysis is developed in a dynamic programming framework, although we are interested in the descriptive modeling of

[2] When an asset market has cleared, there is neither excess demand nor excess supply in the sense that at that moment no market participant is willing to buy the asset at a price equal to or greater than the ask, and no one is willing to sell at a price equal to or less than the bid. Hence we must have a market spread that is at least equal to the minimum allowable price change for the asset in question (on the major U.S. stock exchanges, this is 1/8 of a dollar for most common stocks). However, it is not obvious why spreads greater than minimum allowable price changes are commonly observed.

[3] While the analysis presented here is applicable to any asset market, our formal model treats the secondary market for financial securities. Security markets have two convenient properties: All units of an asset are identical, and such markets are impersonal (which means that bargaining, as distinct from trading, strategies need not be considered).

[4] See n. 2 above.

an investor's decision process rather than in actually generating a normative solution. Section V demonstrates that implementation of the strategic order placement decision (which implies the gravitational pull effect) causes a noninfinitesimal bid-ask spread to exist. This section also defines an equilibrium bid-ask spread, discusses conditions under which it will exist, and shows that it is positively related to market thinness. Section VI considers implications for the design of a security market trading system and summarizes our analysis.

II. The Scenario

Consider an investor who maximizes the expected utility of terminal wealth and, for simplicity, let him allocate funds between only two assets: a risk asset and cash (which we take to be the numeraire asset). In the absence of transaction costs, the market would be monitored continuously and appropriate transactions would be made with each change in the market price and the investor's demand propensities. Then, if price were continuous, there would be no spread and the market price would be determined by a straightforward aggregation of individual demand propensities.

However, a variety of transaction costs impact on the investor's trading decisions. The fixed (with respect to number of shares traded) costs of assessing information, monitoring the market, and conveying orders to the market imply that the investor will make trading decisions only periodically. Further, when decisions are made, he will not convey his full set of demand propensities to the market. For one thing, trades that involve sufficiently small portfolio adjustments would not justify the transaction costs incurred.[5] Also, attempts to transmit several limit orders simultaneously would be likely to overload our current system. Furthermore, a continuous auction which does not generate a Walrasian solution cannot readily handle multiple buy-sell orders that, ex ante, are intended to be alternatives.[6]

In light of transaction costs (and also taking account of the timing and magnitude of exogenous cash flows), the investor will establish a discrete set of decision points. In the analysis presented in Section IV, we take the frequency of these points as predetermined. Upon reaching a decision point, an investor can do nothing, or hit an

[5] It can be readily demonstrated that variable as well as fixed transaction costs make sufficiently small portfolio adjustments prohibitively expensive.

[6] That is, if any array of buy and sell orders from one investor is executed sequentially, the desired quantities at each price should be dependent on the exact sequence of purchases and sales followed. However, the investor does not know ex ante which specific sequence will occur. This problem could, of course, be handled (at a cost) by conditioning orders on the sequence of prior transactions (i.e., if the limit order to buy 200 at $50 executes, then sell 100 at $56).

existing limit order with his own market order, or place his own limit order at a "better" price and run the risk of its not executing. A concurrent strategy issue also exists when the investor finds it prohibitively expensive to convey his entire set of demand propensities to the market; for all intents and purposes, he must select the single best alternative to transmit.

III. The Probability of a Limit Order Executing

In this section, we establish the conditions under which the probability of execution does not approach a limit of unity as the price at which the limit order is placed is taken to be infinitesimally close to the counterpart market quote. Under these conditions we obtain a probability jump (at the counterpart market quote) that underlies the gravitational pull effect developed in Section V. By showing that the probability jump can be attributed to the existence of transaction costs, we establish the basic linkage between spreads and transaction costs.

Consider the case where an investor contemplates submitting a limit bid at the price P_t^{LB} at time t, and let price be a continuous variable. A similar argument can be constructed for the case of a limit ask. Assume that if the limit order is unfilled by the next decision point at time $t + 1$ the order will be canceled. Let L be the length of time between decision points t and $t + 1$.

Let P_t^{MA} be the market ask price at time t. Consistent with the random walk version of the efficient markets hypothesis, we make the Markov assumption that each subsequent market ask depends only on the last previous market ask. If we also assume that each change in the natural log of the market ask, Z_i, is a random variable that is independently and identically distributed over time with mean zero and variance var (Z_i), then we can model the market ask price generation process as a compound Poisson stochastic process:[7]

[7] For expositional simplicity we assume that the stochastic processes considered in this section have no drift (that is, their expected value is zero). It should be noted that even though $\ln P_t^{MA}$ is assumed to have no drift, the price series itself, P_t^{MA}, may have drift. For example, when $\ln (P_{t+1}^{MA}/P_t^{MA})$ is normal, with mean zero and standard deviation σ, then P_{t+1}^{MA}/P_t^{MA} is log normal with mean $\exp(\frac{1}{2}\sigma^2)$.

There will be a realization of the random variable Z_i at each point of time when any one of the following events occurs to affect the specific limit order which sets the market ask: (*a*) it is withdrawn; (*b*) it executes against a crossing buy order; or (*c*) it remains on the book but is no longer the market ask since a lower limit sell has been submitted. Note that events of types *a* and *b* necessarily result in a change in the market ask when price is continuous and utility functions are heterogeneous (which implies a zero probability of having two or more orders at a specific price). Also note that only events of type *b* are associated with transactions; hence the number of Z_i which materialize during any interval will usually exceed the number of transactions in that interval.

$$\ln P_t^{MA}(\Delta) = \ln P_t^{MA} + \sum_{i=1}^{N(\Delta)} Z_i, \tag{1}$$

where Δ is the time from the last decision point. When Δ equals L, we have $P_t^{MA}(\Delta) = P_{t+1}^{MA}$; when Δ equals $2L$, we have $P_t^{MA}(\Delta) = P_{t+2}^{MA}$, etc. The number of changes in the market ask that take place in the time interval Δ is $N(\Delta)$. We assume $N(\Delta)$ follows a Poisson process with arrival rate ν.

The next step is to determine the probability of execution (during a time interval of length L) of a limit bid which is submitted at a price P_t^{LB} greater than the current market bid (P_t^{MB}) but less than the current market ask (P_t^{MA}). Clearly, as the limit bid price approaches (from below) the market ask, the probability of execution increases. One might suppose that this probability approaches unity as the limit bid approaches the market ask; however, this need not be the case. In the Appendix we restate formally and prove:

PROPOSITION 1: If P_t^{MA} is generated by the compound Poisson process of equation (1), then no matter how close the limit bid approaches (from below) the market ask, the probability of the limit order executing is less than unity in any time interval of finite length.

Since a market order will always execute with probability one, proposition 1 gives a probability jump at the market ask.

Transaction costs are crucial to the existence of the probability jump. In the absence of transaction costs, one might expect that, following the work of Merton (1973), the logarithm of the market ask would best be described by a Wiener process with zero drift:

$$d \ln P_t^{MA}(\Delta) = \sigma dZ(\Delta), \tag{2}$$

where σ is the instantaneous variance of the process and $dZ(\Delta)$ is a standardized Wiener process. In this case, the price $P_t^{MA}(\Delta)$ would experience an infinite number of adjustments in the interval $0 \leq \Delta \leq L$. In the Appendix we restate formally and prove:

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then as the limit bid approaches (from below) the market ask, the probability of the limit order executing approaches unity for all time intervals of the finite length.

Proposition 2 implies that for the Wiener process there will be no probability jump at the market ask.

We next consider whether there is a relationship between the compound Poisson process of equation (1) and the Wiener process of equation (2). In the Appendix we restate formally and prove:

PROPOSITION 3: If the random variable Z_i in the compound Poisson process of equation (1) has two equally likely possible values, $+\alpha$ and $-\alpha$, and if the arrival rate ν of this compound Poisson process increases without bound while α simultaneously decreases in such a way that $\nu\alpha^2$ is constant, then the compound Poisson process approaches the Wiener process described by equation (2) with the variance of the Wiener process, σ^2, equaling $\nu\alpha^2$.

Proposition 3 states that the compound Poisson process can be expected to approach the Wiener process under appropriate assumptions.[8] Thus the probability jump of proposition 1 would disappear as ν increased and var (Z_i) decreased. In the Appendix we restate formally and prove:

PROPOSITION 4: As the arrival rate ν in the compound Poisson process of equation (1) increases, the probability of a limit order executing increases for all $P_t^{LB} < P_t^{MA}$.

Hence the probability of a limit order executing increases at all $P_t^{LB} < P_t^{MA}$ as the activity proxy ν increases. Stated conversely, the "thinner" a security, that is, the less active are investors in submitting orders to trade in the security, the lower will be the probability of execution at each $P_t^{LB} < P_t^{MA}$, and therefore the greater will be the probability jump at P_t^{MA}.

The four propositions stated in this section (and proved in the Appendix) have important implications for the analysis of the impact of transaction costs on bid-ask spreads. Without transaction costs, the market price could be expected to behave as a Wiener process; that is, there would be an infinite number of infinitesimally small price adjustments. The investor who was considering placing a limit order could reduce the probability of his order not executing to as close to zero as desired simply by placing his order close enough to the counterpart market quotation.

On the other hand, with transaction costs and a finite number of investors, the market price would generally not behave as a Wiener process. Investors would find continuous adjustments in their portfolios too expensive, and market prices would behave as a stochastic jump process (proposition 3 shows that this process could generally be expected to approach a Wiener process as the order arrival rate is increased). One such jump process is the compound Poisson, and it also is consistent with a (martingale) efficient market.

[8] The particular distribution chosen for Z_i in proposition 3 is not critical. Other appropriately chosen discrete or continuous distributions can also be shown in the limit to go to the Wiener process.

Proposition 1 demonstrates that for such a stochastic jump process a probability jump exists. Proposition 4 has shown that the probability jump will be greater for thinner securities. In Section V we will show that the larger probability jump for thinner issues leads to equilibrium market spreads which are larger for thinner issues.

IV. A Model of Investor Order Placement Strategy

We now consider the question of when an investor will choose to trade via a market order or limit order, or not seek to trade at all. The problem is structured as follows. Because of the costs of monitoring the market, let the investor consider rebalancing his portfolio only at preselected points in time, $t = 1, 2, \ldots, T - 1$, where T is the investment horizon.[9] In order to simplify the analysis, we now consider the placement of only purchase orders for the risk asset (omitting the symmetric case of sell orders without loss of generality).

At any of the $T - 1$ decision points, the investor is faced with three possible courses of action:

a) Submit a market order to buy shares at the current market ask price of P_t^{MA}.

b) Submit a limit order to buy shares at a limit bid price of $P_t^{LB} < P_t^{MA}$.

c) Do nothing.

In modeling the investor's choice among these three alternatives, we find it convenient to make the following assumptions that simplify the analysis without materially affecting the nature of the conclusions. We assume that all orders are for a fixed number of shares ΔN and that when any market or limit order is executed, it is satisfied fully at the stated price; this avoids the tedium of writing (average) transaction price and transaction costs as functions of the number of shares exchanged and of defining probabilities of partial execution. We assume that unfilled limit orders are canceled prior to the next decision point; this avoids both the need to include additional state variables (the price and quantity of any limit order outstanding) and the need to analyze additional courses of action (submit a market order and leave the old limit order outstanding, or submit a market order and remove the limit order, etc.). Finally, we assume no lags in the transmission of information and orders; this avoids the complexity of dealing with changes in the current market quotes during the time the investor formulates and implements his decision.

[9] Note that the rebalancing points need not be the same for all investors, so that trades can occur at any time when the market is open even if specific traders are not continually in the market.

If option a is chosen, the market order will be executed, and the investor's holdings at time $t + 1$ become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{MA} + C^M$ where C^M is the total cost of transmitting and executing the market order.

If option b is chosen, then one of two events can take place: ($b1$) The limit order is executed. The investor's share holdings then become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{LB} + C^{L1} + C^{L2}$ where C^{L1} is the cost of transmitting a limit order to the market and C^{L2} is the cost of executing a limit order. ($b2$) The limit order does not execute and is canceled prior to the next decision point. The investor's share holdings then remain unchanged ($N_{t+1} = N_t$) and his cash is decreased by the cost of transmitting the limit order, C^{L1}. Option b will be chosen over a if the gain associated with the possibility of trading at a more favorable price outweighs the loss associated with the probability of not trading at all.

The investor must consider four subjective probability distributions in order to make an optimal decision. These are: (1) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t; (2) the probability of a limit bid order submitted at time t executing before time $t + 1$; (3) and (4) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t and further conditional on whether a limit bid submitted at time t either did or else did not execute prior to time $t + 1$. In the Appendix we discuss in more detail these four subjective probability distributions (only three of which are independent).

A dynamic programming model of the investor's choice among options a, b, and c is formulated in the Appendix. The investor is assumed to maximize his expected utility of terminal wealth, max (U_1, U_2, U_3), where U_1 is the expected terminal utility of choosing option a, trading via a market order; U_2 is the expected terminal utility of choosing option b, seeking to trade via a limit order; and U_3 is the expected terminal utility of choosing option c, doing nothing. It is convenient to focus on the utility gain, ΔU_1 or ΔU_2, which results from choosing option a or b rather than option c: $\Delta U_1 = U_1 - U_3$; $\Delta U_2(P_t^{LB}) = U_2(P_t^{LB}) - U_3$ (note that U_2 and ΔU_2 are functions of the limit bid price). Clearly $\Delta U_2(P_t^{LB} = P_t^{MA}) = \Delta U_1$, since, at this price, the market order and limit order strategies are effectively the same (the probability is unity of the limit order executing at a price of P_t^{MA}).

Let us now consider the conditions under which $U_1 > \max(U_2, U_3)$, in which case the investor will submit a market order, or $U_2 > \max(U_1, U_3)$, in which case the investor will submit a limit order. Suppose that at the current market quotes the do-nothing strategy is dominated (i.e., $\max[U_1, U_2] > U_3$). Given our utility gains ΔU_1 and

$\Delta U_2(P_t^{LB})$ and the probability function for limit order execution developed in Section III above, we know the utility gain from placing a market order and can readily obtain an expected utility gain function for the limit order strategy. These are depicted in figure 1; the shape of the function is explained as follows:

1. While the utility of a consummated trade decreases monotonically with P_t^{LB}, the probability of execution increases with P_t^{LB}, with two probability jumps (at the market bid and the market ask). The jump at P_t^{MB} simply reflects the institutional reality that orders placed at prices less than or equal to P_t^{MB} would not have priority over the current market bid, whereas an order placed at a price above P_t^{MB} would be at the top of the limit order book. The jump at P_t^{MA} follows from proposition 1.

2. Since the probability of execution is constant to the right of P_t^{MA}, the expected utility gain has a peak at P_t^{MA}, corresponding to the utility gain of transacting by a market order.

3. The probability of execution increases between P_t^{MB} and P_t^{MA}, with the greatest relative increase just to the right of P_t^{MB}. We expect this large probability increase in this neighborhood because the strategy considerations of other investors might lead them to place limit orders just to the right of P_t^{MB} in order to capture the largest price advantage. Hence the second peak in the expected utility gain function will occur at some point P'', between P_t^{MB} and P_t^{MA}.

4. The probability of execution decreases rapidly in a neighborhood just to the left of P_t^{MB} because of existing limit orders on the book. However, since there would be a clustering of limit orders near

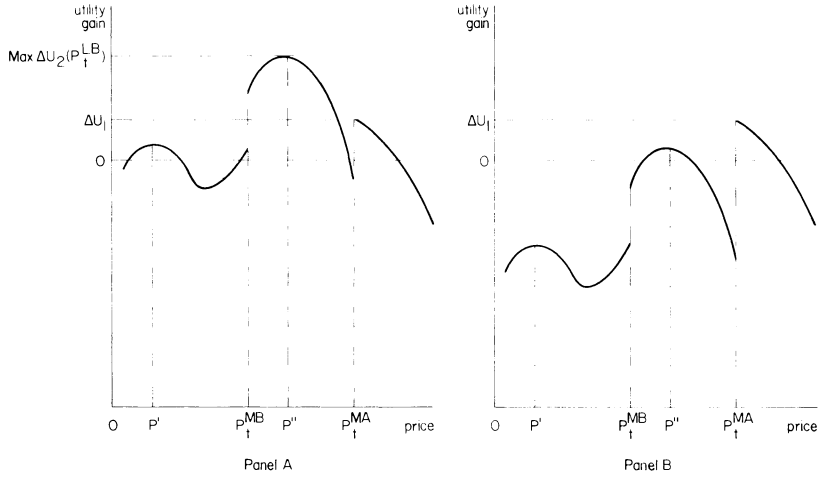

FIG 1.—Illustrative utility gain functions

P_t^{MB}, we would expect the probability of execution to decrease more slowly as the limit price, P_t^{LB}, moves further from P_t^{MB}. Hence, we might find a third peak in the expected utility gain function to occur at some point P' to the left of P_t^{MB}.

Which of the three peaks in $\Delta U_2(P_t^{LB})$ will dominate depends on the particular form of the probability function for limit order execution as well as the expected utility function. In the particular diagram presented in figure 1*A*, $\Delta U_2(P_t^{LB})$ has positive values (hence a do-nothing strategy is dominated), has values greater than ΔU_1 (hence the market order strategy is dominated), and gives a global maximum at P''. Assuming only one limit order per investor, the optimal strategy is to place a limit bid at P'', and hence P'' will become the new market bid. The setting of such a bid quotation can be considered a Stigler (1961)–type search activity.[10]

Clearly the limit order strategy need not always be superior; in selecting this strategy, the investor also accepts the chance that the limit order may fail to execute, in which case the investor would be worse off than had he done nothing since he would have also lost the cost of placing the limit order, C^{L1}. If the probability of failing to execute is high enough and $\Delta U_1 > 0$, the investor will prefer the market order strategy as illustrated in figure 1*B*.[11]

V. The Market Spread

It is clear from the preceding analysis (Sec. IV) that, with a continuous auction market, each investor's order placement decision is made with reference to prices already established on the market—the bid and ask quotations which define the market spread. In turn, each investor's order may affect the market spread to which subsequent traders react. Hence, the market spread is the product of a dynamic interactive process. In this section we show that a nonzero market spread must exist, define the equilibrium market spread, and relate a security's equilibrium spread to its thinness.

The limit order book comprises the limit orders transmitted to the market by a subset of the many traders in a security, and the spread is essentially a gap in the limit order book. Therefore, having estab-

[10] By the placement of such a limit order, a seller or buyer announces his propensity to trade with the hope of getting the attention of a counterpart market participant who would also be willing to trade.

[11] There are two other versions of fig. 1. One would be analogous to panel *A*, but with the global maximum at P' rather than at P''; in this case, the investor would submit a limit bid at P' (below the market bid). The second version would be where the zero point on the utility gain axis is higher than both max $\Delta U_2(P_t^{LB})$ and ΔU_1; in this case, the investor would do nothing.

lished that individual investors will sometimes seek to trade via limit orders, we must show why a noninfinitesimal gap in the array of such orders exists between the market quotes. We do so by considering the impact that the jump in the probability of a limit order executing (Sec. III) has on the investor's optimal order placement strategy (Sec. IV).

We have established the conditions under which a limit order will be placed; we can now consider whether or not limit orders will necessarily be placed so as always to preserve a nonzero bid-ask spread. By continuity of the utility function, and because of the discontinuous drop in the probability of execution that must occur at the market ask as we move to lower prices, we must have a discontinuous drop in $\Delta U_2(P_t^{LB})$ as we move past the market ask to lower prices. Then, because of continuity of the probability function until we get to the market bid (at which point another discontinuous drop occurs), it necessarily follows that $\Delta U_2(P_t^{LB})$ cannot exceed ΔU_1 within a nonzero neighborhood to the left of P_t^{MA}. Hence, no limit bid will be placed infinitesimally close to the market ask, and therefore we must have a nonzero bid-ask spread.

Intuitively viewed, consider an investor who is thinking about whether or not to place a limit bid at some price below the market ask. As the potential limit price rises toward the market ask, it becomes relatively more attractive to transact at the market ask with certainty. At some point, the investor will prefer buying at a discretely higher, unitary probability of execution. In other words, certainty of execution at the market ask creates a kind of gravitational pull which causes investors to jump their price once their potential limit bid gets close enough to the market ask. This clears out limit bids when they get too close to the market ask, thereby leaving a nonzero spread between the market bid and ask.

We next consider the issue of an equilibrium market spread.

DEFINITION: In a dynamic trading process we define the equilibrium market spread as the bid-ask spread at which, for the next instant of time, the probability of the spread increasing is equal to the probability of the spread decreasing.[12]

This definition does not imply that if away from equilibrium the

[12] Note that the probability is not conditioned on the current status of individual investors, recent prices, or other market conditions.

Note that we use the term "equilibrium market spread" to refer to one particular point in the probability distribution of the market spread, rather than to refer to a statistic for the central tendency of the entire distribution. We do not adopt a central tendency definition (such as expected value, mode, or median) of the equilibrium spread because it does not provide a ready link to the market forces that generate the spread and to thinness. For empirical research, the average market spread is likely to be a reasonable proxy for our definition of the equilibrium market spread.

spread will necessarily move toward it on the next order, but only that it is more likely to do so than not.

We assume that a unique equilibrium exists. To examine the reasonableness of this assumption, first consider an arbitrarily small spread. Because the probability of a limit order being placed has a discrete jump at P_t^{MA}, we have already argued that there will be no limit orders placed in some nonzero neighborhood below P_t^{MA} by *all* investors seeking to buy the security. (The generalization from a single investor to all investors takes place simply by choosing the smallest neighborhood found among all investors.) Therefore, for a sufficiently small bid-ask spread, the unconditional probability of the spread increasing must exceed the unconditional probability of the spread decreasing as analyzed from the standpoint of the buyer. A similar argument holds for the seller.

Now consider what occurs when the spread widens. The larger the spread is, the greater the potential utility gain from an optimally priced limit order, while the utility gain from a market order will remain unchanged or fall. The gain in utility occurs because the shares can be bought (sold) at a lower (higher) price via limit order than by market order.[13] With the rise in utility $U_2(P_t^{LB})$, investors will begin to shift their preferences from market orders to limit orders. As more and more investors shift to limit orders, the probability of the spread increasing falls, while the probability of it decreasing rises.

One of two situations must occur. Either the spread would tend to reach a point at which the two probabilities would be equal, in which case it would be the equilibrium market spread, or the limit order book would be empty on one or both sides.[14] Since when the order book is empty market orders cannot transact, we would be at a point where the probability of a limit order is greater than that of a market order. Although such a point might be achieved without the equilibrium market spread ever being attained, we believe it would be atypical.[15]

[13] We assume that the spread does not impart any information to the investor on the direction of future price movements.

[14] No other possibility exists. So long as there is a greater probability that the spread will increase than that it will decrease, the spread will on expectation grow—limited only by the collapse of the limit order book on one side or the other.

[15] With regard to the assumption of uniqueness, investors might have utility functions which could lead to more than one equilibrium market spread. Consider the case where an investor views a very wide spread as signaling the advent of new information which may cause a price adjustment of unknown direction. The investor sensing greater risk might now choose a do-nothing strategy over placing either a limit or market order. This could support a very wide spread as a new market equilibrium spread. Of course, investors would eventually conclude that the new information was not forthcoming, and trading would resume at its former frequency with the old equilibrium spread reestablished.

Our definition of the equilibrium market spread provides a direct link between the market spread and thinness. Recall that proposition 4 showed that the probability of a limit order executing decreases, for all values of P_t^{LB}, as a security's order arrival rate decreases. Thus, ceteris paribus, thin issues have a lower probability of limit orders executing than do thick issues. This lower probability in turn decreases the proportion of investors choosing limit orders over market orders at any given size of the spread; this implies that, for a thinner issue, a wider spread would be required for the forces that increase and decrease the spread to be in equilibrium. Hence we have:

PROPOSITION 5: Thinner securities will, ceteris paribus, have larger equilibrium market spreads.

VI. Conclusion

We have presented a proof of the existence of market spreads in markets with many traders. The formulation treats continuous prices and allows for heterogeneous trading propensities. The proof is not dependent on the demand of traders for immediacy or on the cost to market makers of providing immediacy.

The literature on bid-ask spreads does not appear to have recognized that aggregation from individual to market spreads is a considerably more complex process than the standard aggregation from individual to market demand and supply functions. Neither has it been established that ordinary investors may sometimes seek to trade via limit orders (and at other times via market orders), hence that these investors will sometimes supply (and at other times demand) immediacy, and that in choosing between market and limit orders investors implement an order placement strategy. In addressing these issues, we have sought to establish the links between transaction costs, individual investor order placement strategy, market thinness, and market spreads.

We have first established that, with transaction costs, the probability of a limit order executing does not rise to unity as the price at which the order is placed gets infinitesimally close to a counterpart market quote. We have next shown that the resulting investor trading strategies generate what we have referred to as a gravitational pull effect. Essentially, in the neighborhood of the current market bid and ask quotations, what might otherwise have been limit orders are instead submitted as market orders (at slightly less desirable prices) so as to achieve certainty of execution. These market orders trigger trades which clear limit orders off the book, widening the market spread. The gravitational pull effect explains why market spreads may be

substantial even in markets composed of many traders. Finally, we have defined an equilibrium market spread (where the forces that tend to widen and to narrow the spread are in balance) and have shown it to be positively related to a security's thinness (measured inversely by the order arrival rate).

Our formulation has several implications for the design of a market system. A primary objective of system design should be to expand the extent and frequency with which investors interact with the market by minimizing various transaction costs. Decreasing variable transaction costs will decrease individual spreads and generate a greater order flow; decreasing the costs of monitoring and communicating with the market will also increase the frequency with which investors rebalance their portfolios; and consolidating the currently fragmented system (by, e.g., instituting a consolidated limit order book) will reduce search costs and further shrink spreads by increasing the effective thickness of the market. These costs are all a function of market structure and hence should be amenable to reduction by appropriate system design. However, a major cost of interacting with the market is the cost of decision making, and this might not be subject to significant reduction by exchange organization. For this reason, it is possible that, especially for thinner issues, spreads will remain sizable in a restructured national market system.

Appendix

In this Appendix, we restate formally and prove propositions 1–4 of Section III above and develop the more technical aspects of the arguments presented in Section IV.

Section III

Equation (1) of Section III presents a compound Poisson process as the stochastic process which generates a sequence of market ask prices over time. This can be used to determine the probability that a limit bid order will execute during a particular time period.

Suppose that the price of the limit order to be submitted is greater than the current market bid but less than the current market ask, that is, that $P_t^{MB} < P_t^{LB} < P_t^{MA}$. To determine the probability that the potential limit bid will execute in a time interval of length L, we must find the probability that $P_t^{MA}(\Delta)$ will decrease to a price equal to or less than P_t^{LB}.[16] Therefore, let $Y(L, P_t^{MA}, P_t^{LB})$ be the probability that the minimum value that $P_t^{MA}(\Delta)$ achieves in the interval $0 \leq \Delta \leq L$ is equal to or less than P_t^{LB}. As P_t^{LB} approaches P_t^{MA}, we would expect the probability of the potential limit bid executing to increase, since the amount that $P_t^{MA}(\Delta)$ would have to decline would be reduced. However, in the

[16] For P_t^{MA} discretely greater than P_t^{LB}, the possibility of another investor submitting a limit order with a price greater than P_t^{LB} will only tend to decrease the probability of execution of the original limit order. Thus our proof of existence of the probability jump at P_t^{MA} is conservative.

limit as P_t^{LB} approaches P_t^{MA} from below, without further analysis it is unclear how far the probability will rise. Therefore, let

$$\phi(L, P_t^{MA}) = 1 - \lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}),$$

where the limit is understood to be from below. We now prove:

PROPOSITION 1: If $P_t^{MA}(\Delta)$ is generated by the compound Poisson process of equation (1), then $\phi(L, P_t^{MA}) > 0$ for all intervals of length $L < \infty$.[17]

PROOF: Since Z_i is stochastic with mean zero, $P(Z_i \geq 0) > 0$. Furthermore, the value of the Poisson random variable $N(L)$ will be finite in any interval of length L less than infinity. Therefore, the probability of $N(L)$ consecutive Z_i observations that are greater than or equal to zero is given by $[P(Z_i \geq 0)]^{N(L)}$, which, since $N(L)$ is finite, must be strictly greater than zero. Notice that if all the Z_i observations are greater than or equal to zero, the value of $P_t^{MA}(\Delta)$ must be greater than or equal to P_t^{MA} throughout the interval $0 \leq \Delta \leq L$; thus the limit order P_t^{LB} would not have executed. This is sufficient to demonstrate the probability jump. Clearly, there are other sample paths that $P_t^{MA}(\Delta)$ could have followed which also would have failed to execute the limit order.

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then $\phi(L, P_t^{MA}) = 0$ for all intervals of length $L < \infty$.

PROOF: By the reflection principle for a continuous Wiener process (see Karlin 1968, pp. 276–77) and since $\ln P_t^{MA}(\Delta)$ is driftless,

$$\begin{aligned} Y(L, P_t^{MA}, P_t^{LB}) &= \Pr\{\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) \leq P_t^{LB}\} = \Pr\{\min_{0 \leq \Delta \leq L} \ln P_t^{MA}(\Delta) \leq \ln P_t^{LB}\} \\ &= 2 \Pr\{\ln P_{t+1}^{MA} < \ln P_t^{LB}\} = 2 \Pr\{P_{t+1}^{MA} < P_t^{LB}\} \\ &= \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{LB}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx, \end{aligned}$$

where the latter probability distribution follows from the definition of the Wiener process. Now

$$\lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}) = \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{MA}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx.$$

Substituting the variable $y = (x - P_t^{MA})/(\sigma\sqrt{L})$, the preceding limit equals $(2/\sqrt{2\pi}) \int_{-\infty}^{0} \exp\{-\frac{1}{2} y^2\} dy = 2(\frac{1}{2}) = 1$.

PROPOSITION 3: If the random variable Z_i is expressed as a Bernoulli random variable, with $Pr(Z_i = \alpha) = Pr(Z_i = -\alpha) = 1/2$, and if the arrival rate ν of the Poisson process $N(\Delta)$ goes to infinity, while simultaneously reducing the size of α in such a way that $\nu\alpha^2$ remains constant, the compound Poisson process approaches the Wiener process described by equation (2) with $\sigma^2 = \nu\alpha^2$.

PROOF: This follows from the theorem that the characteristic function for the compound Poisson approaches that of the Wiener process (see Parzen 1962, p. 99).

PROPOSITION 4: $[\partial Y(L, P_t^{MA}, P_t^{LB})]/\partial\nu > 0$ for all $P_t^{LB} < P_t^{MA}$ and $L < \infty$.

PROOF: By the Markov property of the compound Poisson process, increasing (decreasing) the order arrival rate by some factor λ is identical to

[17] This proposition would also hold for other stochastic jump processes where the number of price changes is finite in any finite interval.

increasing (decreasing) the length of time L between decision points by λ. When λ is increased, those sample paths of the process which initially satisfied

$$\min_{0 \leq \Delta \leq L} \quad P_t^{MA}(\Delta) \leq P_t^{LB}$$

will still continue to do so. On the other hand, some sample paths where

$$\min_{0 \leq \Delta \leq L} \quad P_t^{MA}(\Delta) > P_t^{LB}$$

will now satisfy the inequality

$$\min_{0 \leq \Delta \leq \lambda L} \quad P_t^{MA}(\Delta) \leq P_t^{LB}.$$

Therefore, $Y(L, P_t^{MA}, P_t^{LB})$ would increase for $\lambda > 1$ for all values of $P_t^{LB} < P_t^{MA}$. Similarly, $Y(L, P_t^{MA}, P_t^{LB})$ would decrease if the order arrival rate λ is decreased.

Section IV

We now present the more technical aspects of the model of an investor's order placement strategy discussed in Section IV. The subjectively determined probability distributions are specified as follows. First, the probability distribution of future market bid and ask prices is given by $h(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB})$, where h is a joint density function for market asks and bids $(P_{t+1}^{MA}, P_{t+1}^{MB})$ at time $t + 1$, given the prices at t. Consistent with random walk models of security price behavior, we condition future prices only on current prices. Consistent with Section III, we let the investor's subjective probability of a limit order executing before $t + 1$ be given by $p(P_t^{LB}, P_t^{MA}, P_t^{MB})$, where P_t^{LB} is the limit bid price. (Again, this is consistent with random walk theory, since we need know only current market quotes to predict whether or not the limit order will be executed.) Last, we assume the investor determines conditional probability distributions of future prices in the event of either a successful or unsuccessful limit order. That is, for a limit bid at price P_t^{LB} submitted at time t, let $k(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function of market bid and ask prices at time $t + 1$ if the limit bid executes prior to $t + 1$, and let $l(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function at time $t + 1$ if the limit bid fails to execute prior to $t + 1$. Note that only three of the four subjective probability distributions h, p, k, and l can be independently determined, since by Bayes's theorem we must have

$$\begin{aligned} h(x, y \mid P_t^{MA}, P_t^{MB}) = {} & p(P_t^{LB}, P_t^{MA}, P_t^{MB}) k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \\ & + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \end{aligned}$$

for all possible values of P_t^{MA}, P_t^{MB}, and P_t^{LB}.

We can now give a formal statement of the model. Given observed market quotes of P_t^{MA} and P_t^{MB} and holdings by the investor of N_t shares of the security and S_t dollars in cash, at time t the investor's expected utility of terminal wealth can be written as $\psi_t = f_t(P_t^{MA}, P_t^{MB}, N_t, S_t)$.[18]

[18] Note that the investor's underlying utility function measures the utility of the wealth he will possess at some horizon T, where $T > t$. The investor will choose those decisions at times $t, t + 1, \ldots, T - 1$ which maximize the expected utility of his terminal wealth. As of time t (the investor's current decision point), viewing the expectations operator as ranging over relevant random variables pertaining to times between t and T, we define ψ_t to be the expected utility of the investor's terminal wealth. Clearly this depends upon the assets the investor has at time t (N_t and S_t) and the current market prices at which the investor can buy or sell shares (P_t^{MA} and P_t^{MB}).

We now derive an expression for ψ_t in terms of the various probability assessments and costs. Write $\psi_t = \max(U_1, U_2, U_3)$, where U_1 = expected terminal utility if a market order is placed at time t, U_2 = expected terminal utility if a limit order is placed at time t, and U_3 = expected terminal utility of doing nothing at time t. More precisely, letting x and y be values for the ask and bid prices at time $t + 1$, respectively, we have

$$U_1 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t + \Delta N, S_t - C^M - N \cdot P_t^{MA}) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dx dy.$$

As stated in Section IV, we assume that the number of shares to be purchased, ΔN, is fixed. Define the function

$$\begin{aligned} U_2(P_t^{LB}) = {} & p(P_t^{LB}, P_t^{MA}, P_t^{MB}) \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t + \Delta N, \\ & S_t - C^{L1} - C^{L2} - \Delta N \cdot P_t^{LB}) \cdot k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dx dy \\ & + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t - C^{L1}) \\ & \cdot l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dx dy \end{aligned}$$

for all possible prices of the limit order. Then we have

$$U_2 = \max_{P_t^{LB}} U_2(P_t^{LB}).$$

Finally, for the do-nothing option we have

$$U_3 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dx dy.$$

The dynamic programming recursion permits us to obtain a solution for f_t in the order $T - 1, T - 2, \ldots, 2, 1$. This is possible since we assume the utility value for f_T is known for all values of the parameter P_T^{MA}, P_T^{MB}, N_T, and S_T. We have defined this recursion in terms of the four state variables P_t^{MA}, P_t^{MB}, N_t, and S_t. We also have two decision variables, P_t^{LB} and the decision as to which of the three courses of action to take.

References

Amihud, Yakov, and Mendelson, Haim. "Dealership Market: Market-making with Inventory." *J. Financial Econ.* 8 (March 1980): 31–53.

Beja, Avraham, and Hakansson, Nils H. "Dynamic Market Processes and the Rewards to Up-to-Date Information." *J. Finance* 32 (May 1977): 291–304.

Benston, George J., and Hagerman, Robert L. "Determinants of Bid-Asked Spreads in the Over-the-Counter Market." *J. Financial Econ.* 1 (December 1974): 353–64.

Branch, Ben, and Freed, Walter. "Bid-Asked Spreads on the Amex and the Big Board." *J. Finance* 32 (March 1977): 159–63.

Cohen, Kalman J.; Hawawini, Gabriel A.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb, David K. "Implications of Microstructure Theory for Empirical Research on Stock Price Behavior." *J. Finance* 35 (May 1980): 249–57.

Cohen, Kalman J.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb,

David K. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Securities Markets." *J. Finance* 33 (March 1978): 149–67.

———. "Market Makers and the Market Spread: A Review of Recent Literature." *J. Financial and Quantitative Analysis* 14 (November 1979): 813–35.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82 (February 1968): 33–53.

Garman, Mark B. "Market Microstructure." *J. Financial Econ.* 3 (June 1976): 257–75.

Hamilton, James L. "Competition, Scale Economies, and Transaction Cost in the Stock Market." *J. Financial and Quantitative Analysis* 11 (December 1976): 779–802.

———. "Marketplace Organization and Marketability: NASDAQ, the Stock Exchange, and the National Market System." *J. Finance* 33 (May 1978): 487–503.

Ho, Thomas, and Stoll, Hans R. "Optimal Dealer Pricing under Transactions and Return Uncertainty." Working Paper, New York Univ., Graduate School Bus. Admin., 1979.

———. "On Dealer Markets under Competition." *J. Finance* 35 (May 1980): 259–67.

Karlin, Samuel. *A First Course in Stochastic Processes.* 2d ed., enl. New York: Academic Press, 1968.

Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (September 1973): 867–87.

Newton, William, and Quandt, Richard E. "An Empirical Study of Spreads." Working Paper, Princeton Univ., Dept. Econ., 1979.

Parzen, Emanuel. *Stochastic Processes.* San Francisco: Holden-Day, 1962.

Schleef, Harald J., and Mildenstein, Eckhard. "A Dynamic Model of the Security Dealer's Bid and Ask Prices." Paper presented at meetings of the Western Economic Association, Las Vegas, 1979.

Smidt, Seymour. "Continuous vs. Intermittent Trading on Auction Markets." Working Paper, Cornell Univ., Graduate School Bus. and Public Admin., 1979.

Stigler, George J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25.

———. "Public Regulation of the Securities Markets." *J. Bus.* 37 (April 1964): 117–42.

Stoll, Hans R. "The Supply of Dealer Services in Securities Markets." *J. Finance* 33 (September 1978): 1133–51. (*a*)

———. "The Pricing of Security Dealer Services: An Empirical Study of NASDAQ Stocks." *J. Finance* 33 (September 1978): 1153–72. (*b*)

Tinic, Seha M. "The Economics of Liquidity Services." *Q.J.E.* 86 (February 1972): 79–93.

Tinic, Seha M., and West, Richard R. "Competition and the Pricing of Dealer Service in the Over-the-Counter Stock Market." *J. Financial and Quantitative Analysis* 7 (June 1972): 1707–27.

———. "Marketability of Common Stocks in Canada and the U.S.A.: A Comparison of Agent versus Dealer Dominated Markets." *J. Finance* 29 (June 1974): 729–46.

U.S. Securities and Exchange Commission. *Institutional Investor Study Report.* 92d Cong., 1st sess. House Document no. 92–64, 1971.

West, Richard R., and Tinic, Seha M. *The Economics of the Stock Market.* New York: Praeger, 1971.