THE JOURNAL OF FINANCE • VOL. XLIII, NO. 3 • JULY 1988

Liquidity and Market Structure

SANFORD J. GROSSMAN and MERTON H. MILLER*

ABSTRACT

Market liquidity is modeled as being determined by the demand and supply of immediacy. Exogenous liquidity events coupled with the risk of delayed trade create a demand for immediacy. Market makers supply immediacy by their continuous presence and willingness to bear risk during the time period between the arrival of final buyers and sellers. In the long run the number of market makers adjusts to equate the supply and demand for immediacy. This determines the equilibrium level of liquidity in the market. The lower is the autocorrelation in rates of return, the higher is the equilibrium level of liquidity.

KEYNES ONCE OBSERVED THAT while most of us could surely agree that Queen Victoria was a happier woman but a less successful monarch than Queen Elizabeth I, we would be hard put to restate that notion in precise mathematical terms. Keynes' observation could apply with equal force to the notion of market liquidity. The T-bond Futures pit at the Chicago Board of Trade is surely more liquid than the local market for residential housing. But how much more? What is the decisive difference between them? Is the colorful open-outcry format of the T-bond Futures market the source of its great liquidity? Or does the causation run the other way?

Those are some of the issues we propose to consider here. Our purpose is to present a simple model of market structure that captures the essence of market liquidity. A key feature of the model is its finer partitioning of time intervals and of roles for market participants than in standard treatments of the determination of market prices. Much economic theory, in the Walrasian tradition, still proceeds as if prices were set in a gigantic town meeting in which all potential buyers and sellers participate directly. Researchers in the rapidly growing specialty, sometimes dubbed market microstructure theory, have expanded the cast to include market makers in the sense of intermediaries who can fill gaps arising from imperfect synchronization between the arrivals of the buyers and the sellers. The focus of this literature has been on the inventory-management policies of market makers (see, e.g., Stoll [13]) and on their responses to the threat of adverse information trading against them (see, e.g., Glosten and Milgrom [5]). Our intention here, however, is not to expand this important and interesting class of inventory models but to fit these intermediaries and their temporary inventory holdings into a larger framework that also encompasses the ultimate demanders and suppliers.

* Department of Economics, Princeton University, and Graduate School of Business, University of Chicago, respectively. This paper was presented at the Annual Meetings of the American Finance Association, December 29, 1987, in Chicago. Helpful comments on an earlier draft were received from Kenneth Cone, Kenneth French, T. Eric Kilcollin, Andrei Shleifer, Lester Telser, and Robert Vishny.

A Brief Overview of the Model: The Supply and Demand for Immediacy

Our model of market structure has two participant groups, and we shall refer to them, for simplicity, as *market makers*, and *outside customers*. For simplicity of exposition only, we shall take their basic tastes, including risk tolerances, as the same. Their roles are defined at this stage principally in terms of their initial endowments.

Within the group of outside customers are some who, for any of a variety of reasons, experience what we call a *liquidity event*, which leads them to perceive a gap at current prices between their desired holdings of a particular asset and their current holdings of that asset. Even if the gaps sum to zero across the whole group, as we assume, some customers might propose to remedy their portfolio imbalance immediately by undertaking a transaction in the asset; for concreteness in exposition, suppose these potential liquidity traders are net sellers. In our model the putative sellers can choose to offer the goods immediately to the market makers who happen to be in the market currently and who have no holdings of the asset, or at least no imbalance that they too are seeking to eliminate. Or, a seller can postpone the offer to sell for one stylized period until the potential buyer customers on the other side of the trade have learned of the offer and have had a chance to come to the market.

Clearly the seller faces a trade-off. By waiting until more potential buyers have been notified, the seller increases the chance of finding an eager buyer. But this delay carries risks; while the buyers are assembling, the ultimate equilibrium price may shift. The best selling price for a sale delayed to the second period may be substantially lower (or higher) than the price in a sale to a market maker in the first period. By selling immediately, that interim price risk is transferred to the market maker who then waits until the ultimate buyers have assembled. When we speak of *the demand for immediacy* by a seller, we mean the willingness to sell rather than wait. This demand depends on the volatility of the underlying price and the diversifiability of the risk of an adverse price move.

The market makers charge for bearing price risk by offering the immediate sellers a price that is not uncertain, but that is lower, on average, than the sellers could expect from delaying. The expected price rise between periods 1 and 2 is, of course, only the market maker's gross return before allowing for the costs of supplying the service. These costs include not only any direct costs of effecting and monitoring trades, but also the important, though often overlooked cost of being available and open for business when the outside customers arrive to trade. These opportunity costs of maintaining a continuous presence in the market, which we model as fixed costs, play a key role in determining *the supply of immediacy* and market-making services.

The market makers, as emphasized earlier, must also assume the price risk that the immediacy demanders shed. That the aggregate price risk is merely shifted to the market makers does not, however, rule out efficiency gains from the arrangement. In our model, where all participants have the same risk tolerance, the gains arise essentially from diversification—the spreading of the transferred risks over the entire group of market makers. The larger that group, the lower, *ceteris paribus*, the risk and expected return per unit traded by each

and hence also the lower the effective cost of immediacy to the customers. The number of market makers will adjust until, in equilibrium, the returns to each from assuming the risk of waiting to trade with the ultimate buyers just balance the costs of maintaining a continuous presence in the market. This adjustment determines the equilibrium amount of immediacy provided, i.e., the amount by which price is temporarily depressed by a typical sell order.

Our model suggests looking to differences in the cost to market makers of maintaining a market presence and to differences in the demand by customers for immediacy for the keys to market structure and market liquidity. The greater the demand for immediacy and the lower the cost to market makers of maintaining a continuous presence, the larger the proportion of the transactions between ultimate customers effected initially through market makers, and hence the more liquid the market.

The Liquidity Spectrum in Real-World Market Structures

Successful futures markets are the leading examples of markets where the demand for immediacy is high. Futures markets are successful precisely for those commodities and in those time periods where price volatility, and hence the risks of delaying trading, is high. The price risks of volatility are further reinforced for potential hedger customers in those markets by the high leverage and extreme under-diversification of the underlying spot inventory positions that constitute their main line of business. Immediacy also becomes of particular concern where, as is frequently the case, the futures transaction is merely one leg of an inter-contract or inter-market hedge. Little or no risk may be incurred once all the components of the hedge have been put in place, but much risk is incurred when only some of the legs have been set. When the transactor is "naked," to use the colorful language of the trade, the delay of even a few seconds can become critical. (See, e.g., Grossman and Miller [6].)

The demand for immediacy in successful futures markets is not only urgent, but sustained. The regular seasonal build up and build down of inventories, as commodities move through the production chain, creates a continual desire to *trade*, not just to *hold* futures. In financial futures markets, dealers' inventories of the underlying securities build up and down in response to periodic auctions of U.S. Treasury issues, to the flotation of stocks or bonds by corporations, or to the restructuring of portfolios by large institutional investors.

The sustained demand for hedging and hence for trading futures quickly is often accommodated by designating a specific physical market place or exchange in which many competing market makers can offer their services simultaneously. Such arrangements help spread the fixed costs to market makers of maintaining a presence, as does the practice at most present-day futures exchanges of providing trading areas for many different contracts between which individual market makers can drift as trading interest changes. Many, but not all, futures exchanges also permit market makers to serve both as brokers for customers and as traders on personal account, though not, of course, on the same transaction. Most floor traders tend to specialize in one role or the other, but the freedom to

switch roles can permit a quick adjustment in the number of market makers when the flow of orders changes abruptly.[1]

At the opposite extreme from the highly liquid futures markets, where intermediary market makers participate as principals in virtually all transactions, stand the highly illiquid markets, such as those for residential housing, where virtually none of the transactions pass through a dealer's temporary inventory.[2] Sellers of individual homes are typically less concerned with short-term price volatility, and hence with immediacy, than with making sure that the widest possible set of ultimate buyers can be informed of the house's availability. Potential market makers, moreover, face not only all the ordinary costs of maintaining a continuous presence in a thin market, but the additional moral hazards that arise from the owner's possibly adverse private information about the value of the property. The result is a market in which intermediaries, to the extent that they are involved at all, provide brokerage or search services, not immediacy.

The Structure of the Stock Market

Most real-world markets lie somewhere between these liquidity extremes and their structures will typically mix features from both the search markets and the liquidity markets. U.S. stock market institutions, for example, currently involve at least four distinct forms of market organization operating simultaneously, but in different segments of the market and with somewhat different immediacy clienteles:

1. For a few of the most widely held and heavily traded securities, such as IBM or AT&T, the market at the New York Stock Exchange often approximates the open-outcry pits at the commodity exchanges. These are stocks in which the minute-to-minute order flow is highly variable relative to the arrival of news about the underlying value of the shares, and for which our model predicts a large number of market makers in equilibrium. The "crowd" for those stocks, though substantially smaller than in the T-bond futures market, is large enough to offer a competitive discipline to the Exchange's franchised "specialist," who, in these particularly active markets, typically plays more the role of an auctioneer (and a commission collector) than a market maker on personal account.

2. The specialist's role as a market maker assumes greater prominence for the hundreds of smaller, less active stocks, some of which may not even trade as frequently as once a day. In such stocks, our model would not predict an equilibrium with many market makers. The designation of a specialist by the

[1] For a discussion of the benefits and the supposed abuses of dual trading on futures exchanges, see Grossman and Miller [6].

[2] Although the fraction of potential trades executed immediately by market makers rather than delayed for search is higher for futures exchanges than in virtually any other market setting, search plays a role even there. A case in point is so-called "sunshine trading" in which pending large and presumably informationless orders by portfolio insurers are publicized in advance throughout the investment community with a view to attracting a large inflow of counterparties prepared to take the other side. Whether such sunshine trading violates long-standing regulatory prohibitions against "prearranged trading" is a policy issue currently much in dispute.

Exchange, however, does at least guarantee that someone will indeed be maintaining a physical presence in the market, ready to effect a transaction should an order happen to arrive. The potential for abuse of the specialist's monopoly position is mitigated by the same standard cross-subsidization approach long familiar in U.S. public-utility regulation. As a condition for keeping the franchise, specialists on the New York Stock Exchange, for example, are encouraged by the Exchange to limit price changes between successive transactions to no more than one tick (normally 12½ cents per share), using personal inventory to absorb any temporary imbalances along the way. This restriction, which is in fact monitored by the Exchange, serves both to limit specialists' profit and to create the appearance of liquidity, though in practice only for very small transactions. Should a very large order arrive, however, and should it be larger than can be absorbed by the specialist or by any previously entered "limit orders" then resting on the specialist's "book," the market can switch to search mode. The specialist, with the permission of the Exchange, can suspend trading in the stock and institute a search for counterparties to the imbalance, either elsewhere on the floor of the Exchange or, more likely these days, off the floor at the block-trading desks of the investment bankers.

3. These desks are the third, and increasingly the dominant form of market organization for trading common stocks in the U.S., thanks to the concentration of so much corporate stock in a relatively small number of extremely large pension funds, mutual funds and other institutional holders. Because relatively small portfolio adjustments by these institutional holders would be far too large to be absorbed by any specialist firm, the large blocks of single stocks, or sometimes whole portfolios are brought to the "upstairs market" maintained by the investment banking firms. Until recently at least, the upstairs desks functioned primarily as a search market. The upstairs traders essentially "shopped the block" among their customers, and when a suitable counterparty had been located and a deal struck, they reported the trade to the relevant specialists on the floor of the Exchange. In the process, they picked up on behalf of the initiating side any limit orders on the specialist's book that were transformed into market orders by the price change occasioned by the block trade.

Although search was the initial, and still remains the major function of the upstairs market, the amount of "positioning" and hence of market-making liquidity provided by the upstairs firms has increased substantially in recent years. The shift traces mainly to the highly liquid futures and options index markets which permit the upstairs firms to hedge their inventories while conducting the search for or waiting for the other side of the transaction.

4. Finally, at the other end of the spectrum from the upstairs, wholesale broker-dealer market lies the retail, dealer market in Over-the-Counter (OTC) stocks, for which, with a few well-known exceptions, the normal trading interest is typically too small to justify listing even on a regional exchange.[3] The market for such stocks is not a physical exchange floor but a set of computer terminals.

[3] Some corporations of substantial size, however, may nevertheless choose to list in this market because there are fewer restrictions on size and capital structure (such as a one-share, one-vote rule) than on the NYSE or AMEX.

When introduced originally in the 1970's, the computerized NASDAQ market system for OTC stocks offered essentially only a "bulletin board" in which those market makers with access to the system could enter price quotes. The quotes, though deemed firm for some standard, minimum-size trade, were essentially advertisements, and the actual transactions were not executed automatically, but negotiated between the parties. The market makers in particular stocks, although they did position small inventories, assumed no obligation to maintain a continuous presence or to smooth price changes between successive transactions.[4]

All four forms of market organization for trading common stocks, along with those of the index futures and options markets, were subjected in October 1987 to what seemed to be liquidity events, in our sense, of unprecedented magnitude. We shall return briefly to those events in Section II. First, however, we turn in Section I to set down the detailed structure of our model of market liquidity and a characterization of its equilibrium.

I. A Formal Model of Market Liquidity

In this section we present a formal model of the role of market makers in providing immediacy. We focus most of our attention on the consequences of a temporary order imbalance of size i in a simple world with only three dates: 1, 2, and 3. At date 1 a liquidity event occurs that creates a temporary order imbalance of size i. Market makers offset this temporary imbalance by taking trading positions that they hold until date 2. We denote the nonmarket makers by the term "outside customer" although in practice, of course, individuals and firms can play either role at different times. By a temporary order imbalance we mean an asynchronization of outside-customer trading times; the net trading demand would be zero at the current price if all traders were simultaneously present in the market.

At date 2, the market makers offset their positions as other outside customers arrive to offset the imbalance. Thus, the length of time between date 1 and date 2 is the period of time needed for enough orders of outside customers to arrive at the market to offset the initial order imbalance. Date 3 is introduced only as a terminal condition for valuing the securities as of date 2.[5]

We assume two assets: a risk-free asset called cash (with zero rate of return), and a risky asset. Let $\tilde{P}_3$ be the exogenously given terminal price (or liquidation

[4] In addition to the four markets so far listed, there may now be as many as six distinct stock markets if one counts the "after-hours" market (which now includes the trading of big-name U.S. stocks on foreign exchanges) and the so-called "fourth market" in which large pension funds, especially those following "passive" or indexing strategies, transfer baskets of stocks directly to and from each other in essentially informationless trades. The futures and options markets in stocks, of course, constitute still another form of stock market at least for the trading if not the holding of stocks.

Many European stock markets, where the volumes of trading are still quite small by U.S. standards, use "batch" or "periodic call" systems rather than any of the continuous-trading systems we find here. For a comprehensive survey of trading practices overseas, see Whitcomb [14].

[5] The reader is referred to Ho [9] for a model of equilibrium market making in a continuous-time, Poisson-arrival-of-orders setting.

value) of the risky asset. Assume that public information about $\tilde{P}_3$ arrives before trade at period 1 and also before trade at period 2. Let $\tilde{x}_t$ be the number of units of the asset owned by an outside customer after trade at time t, and let B_t be that customer's holdings of cash (in dollars). Two interpretations can be given to $\tilde{x}_3$. In the first, the outside customer is a commercial hedger and the asset is a futures contract. In this case, the hedger's net holding at period 3 is $\tilde{x}_3 = \tilde{x}_2 + i$, where i is the number of units of the spot commodity (which may, of course, be a security) owned by the hedger. The hedger's terminal wealth is then

$$W_3 = B_2 + P_3\tilde{x}_3 = B_2 + P_3\tilde{x}_2 + iP_3. \tag{1}$$

The hedger is using the futures market to offset the spot price $\tilde{P}_3$ risk of the initial position. (Note that under this futures market interpretation, the asset is in zero supply.)

In the second interpretation, the market is a stock market, and the outside customer at time 1 has an endowment of size i in the security, which is inappropriate in the light of the customer's risk preferences and information on the risk-return pattern associated with the security. In this case, $\tilde{x}_3$ is the final holding of the security at the terminal date, and $\tilde{x}_3 = \tilde{x}_2$. In contrast to the futures market interpretation, the asset is not in zero supply, and if i is correlated across customers, then the aggregate endowment of the asset relevant for market clearing at each date t will be affected by i.

Under either interpretation we assume that at times $t = 1, 2$ the customer chooses asset holdings $\tilde{x}_t$ and a risk-free asset position B_t to maximize the expected utility of terminal (i.e., date-3) wealth

$$\mathrm{E}U(W_3)$$

subject to

$$W_3 = B_2 + \tilde{x}_3\tilde{P}_3 \tag{2a}$$

$$\tilde{P}_2\tilde{x}_2 + B_2 = W_2 = B_1 + \tilde{P}_2\tilde{x}_1 \tag{2b}$$

$$P_1\tilde{x}_1 + B_1 = W_1 = P_1 i_1 + W_0, \tag{2c}$$

where i_1 represents the initial endowment of the asset and W_0 represents other wealth. Note that:

$$i_1 = 0 \quad \text{and} \quad \tilde{x}_3 = \tilde{x}_2 + i \text{ in the futures market case;}$$

$$i_1 = i \quad \text{and} \quad \tilde{x}_3 = \tilde{x}_2 \text{ in the stock market case.}$$

If B_1 and B_2 are eliminated from (2a)–(2c) we obtain

$$W_3 = W_0 + (\tilde{P}_2 - \tilde{P}_1)(\tilde{x}_1 - i_1) + (\tilde{P}_3 - \tilde{P}_2)(\tilde{x}_2 - i_1) + \tilde{P}_3 i,$$

where $\tilde{x}_t - i_1$ represents the excess demand for the asset, whether it be a futures contract or a stock. Therefore, it simplifies matters to define a trader's *excess* demand to be

$$x_t = \tilde{x}_t - i_1 \qquad t = 1, 2.$$

In the above notation customers choose their positions to maximize

$$\mathrm{E}U(W_3) = \mathrm{E}U(W_0 + (\tilde{P}_2 - P_1)x_1 + (\tilde{P}_3 - \tilde{P}_2)x_2 + \tilde{P}_3 i). \tag{3}$$

We will assume that $\tilde{P}$ is normally distributed at each date, and that

$$U(W) = -e^{-aW}. \tag{4}$$

By backward induction and (4), if we let x_2^{cd} denote the optimal value of x_2 (chosen at date 2), then x_2^{cd} solves

$$\max_{x_2} \mathrm{E}_2 U(W_2 - P_2 i_1 + (\tilde{P}_3 - P_2)x_2 + \tilde{P}_3 i).$$

Using the exponential utility function, the optimal value for x_2 is

$$x_2^{cd} = \frac{\mathrm{E}_2\tilde{P}_3 - P_2}{a\ \mathrm{Var}_2\tilde{P}_3} - i, \tag{5}$$

where all means and variances are conditioned on the information at time 2. Note that the customer's excess demand is x_2^{cd} which is linear in i. Hence, if all customers are identical, except possibly with respect to i, we can take x_2^{cd} to represent the aggregate demand of customers, and i to be the aggregate potential imbalance.

We assume that there are M other traders in the market who do not hold the spot commodity and thus face no spot-price risk; these are the market makers. (Under the stock market interpretation, assume that the market makers do not hold an endowment of the security prior to their date-1 trading with outside customers.) Market makers have the same utility function, but for them $i = 0$. Hence, if the excess demand per market maker is x_2^{md}, the total excess demand by market makers in period 2 is

$$Mx_2^{md} = M\,\frac{\mathrm{E}_2\tilde{P}_3 - P_2}{a\ \mathrm{Var}_2\tilde{P}_3}. \tag{6}$$

We now state the assumption critical to understanding the benefits of waiting from period 1 to period 2 to trade. In particular, it is that *asynchronization* of desired trades creates the demand for immediacy at time 1. Thus, the positive immediacy demand felt by the customers at time 1 is, by definition, offset by *new* customers arriving at date 2 with the opposite imbalance from those who arrived at date 1. Their aggregate excess demand is

$$\frac{\mathrm{E}_2\tilde{P}_3 - P_2}{a\ \mathrm{Var}_2\tilde{P}_3} + i.$$

Market clearing at date 2 requires that the excess demand of (a) customers who arrived at date 1, plus (b) market makers, plus (c) the new customers arriving at date 2 should sum to zero:

$$\frac{\mathrm{E}_2\tilde{P}_3 - P_2}{a\ \mathrm{Var}_2\tilde{P}_3} - i + M\,\frac{(\mathrm{E}_2\tilde{P}_3 - P_2)}{a\ \mathrm{Var}_2\tilde{P}_3} + \frac{\mathrm{E}_2\tilde{P}_3 - P_2}{a\ \mathrm{Var}_2\tilde{P}_3} + i = 0. \tag{7}$$

Note that in a futures interpretation, the right-hand side represents an aggregate endowment of zero, while in the stock market (7) means that excess demands

(i.e., trades net of endowments) must sum to zero. Note also that under our convention that period 3 is merely a terminal condition, (7) implies:

$$\mathrm{E}_2\tilde{P}_3 - P_2 = 0. \tag{8}$$

The equilibrium excess demand at date 2 of the customer arriving at the market at date 1 is thus

$$x_2^{cd} = -i. \tag{9}$$

Using (3), (4), (8), and (9) we can find the date-1 demand of the customer from

$$\underset{x_1}{\mathrm{Max}}\ \mathrm{E}_1 U(W_0 + x_1(\mathrm{E}_2\tilde{P}_3 - P_1) + i\mathrm{E}_2\tilde{P}_3). \tag{10}$$

This problem has the same form as the problem in period 2 except that the risk from the point of view of period 1 is that $P_2 = \mathrm{E}_2\tilde{P}_3$ is not known. As before, the customer's excess-demand function is

$$x_1^{cd} = \frac{\mathrm{E}_1\tilde{P}_3 - P_1}{a\ \mathrm{Var}_1(\mathrm{E}_2\tilde{P}_3)} - i, \tag{11}$$

where the law of iterated expectations is used to obtain $\mathrm{E}_1\mathrm{E}_2\tilde{P}_3 = \mathrm{E}_1\tilde{P}_3$.

A. Market Makers and the Provision of Immediacy

At date 1, there are M market makers. They constantly watch the floor of the exchange either directly or through their agents on the floor. They solve the same maximization problem as the customers except that for them $i = 0$. Hence their excess-demand function is

$$x_1^m = \frac{\mathrm{E}_1\tilde{P}_3 - P_1}{a\ \mathrm{Var}_1(\mathrm{E}_2P_3)}. \tag{12}$$

Market clearing at date 1 thus requires

$$Mx_1^m + x_1^{cd} = 0. \tag{13}$$

Using (11) and (12), it is seen that (13) becomes:

$$\frac{\mathrm{E}_1\tilde{P}_3 - P_1}{a\ \mathrm{Var}_1(\mathrm{E}_2\tilde{P}_3)} = \frac{i}{1 + M}. \tag{14}$$

Let $\tilde{r} = \tilde{P}_2/P_1 - 1$ be the excess return earned by the market makers. Then

$$\mathrm{E}_1\tilde{r} = \frac{P_1 i}{1 + M}\, a\ \mathrm{Var}_1(\tilde{r}). \tag{15}$$

Thus if M is finite, a positive value of $P_1 i$ (which causes hedgers to desire to "short") will induce a temporary fall in the market price. Note that we have defined the order imbalance to sum to zero across periods 1 and 2. In particular, no aggregate risk is associated with holding the asset across periods. Therefore, in the absence of an asynchronization of order flows, $\mathrm{E}_1\tilde{r} = 0$. It is the asynchronization of these flows and the finite risk-bearing capacity of market makers that

leads $E_1\tilde{r}$ to deviate from 0. Note that from (12) and (14) the value of the positions held by a typical market maker (i.e., his or her inventory) is

$$P_1 x_1^m = \frac{P_1 i}{1 + M}.$$

The larger is this inventory the higher the expected return between period 1 and 2 to compensate the market maker for the risk that new information may arrive (causing $E_2\tilde{P}_3 = P_2 \neq P_1$) leading to capital losses on the inventory positions.

B. *Determination of the Number of Market Makers*

A market maker choosing always to have a presence on the trading floor is assumed to forego opportunities elsewhere worth \$c. We assume that the size and direction of the liquidity event i is not known at the time that this cost is "sunk." We represent i as the realization of a normally distributed random variable, uncorrelated with information about $\tilde{P}_3$. The gain from being on the floor is the ability to trade at price P_1. Then the expected utility of a market maker who pays \$c out of initial wealth is

$$EU(W_0 - c + (\tilde{P}_2 - \tilde{P}_1)x_1^m),$$

where the profit between period 2 and period 3, $(\tilde{P}_3 - P_2)x_2^m$, does not appear because (6) and (8) imply that $x_2^m = 0$.

Free entry of market makers will occur until

$$EU(W_0 - c + (\tilde{P}_2 - P_1)x_1^m) = EU(W_0). \tag{16}$$

Equation (12) and the exponential utility assumption can be used to evaluate (16):

$$e^{ac}\mathrm{E}\exp\left(-\left(\frac{a^2}{2}\right)(\mathrm{Var}_1\tilde{P}_2)\left(\frac{i}{1 + M}\right)^2\right) = 1 \tag{17a}$$

or

$$e^{ac}\mathrm{E}\exp\left(-\frac{t}{2}z^2\right) = 1, \tag{17b}$$

where

$$t = a^2\frac{\mathrm{Var}_1\tilde{P}_2}{(1 + M)^2}\,\mathrm{Var}\,i, \quad \text{and} \quad z^2 = \frac{i^2}{\mathrm{Var}\,i}.$$

Using the moment-generating function of the non-central Chi-squared distribution, (17b) becomes

$$\frac{1}{\sqrt{1 + t}}\exp\left(\frac{-(\mathrm{E}i)^2}{1 + t}\left(\frac{t}{2}\right)\right) = e^{-ac}. \tag{18}$$

If we assume that the expectation of an order imbalance is zero, i.e., $Ei = 0$, then (18) becomes

$$\frac{1}{\sqrt{1+t}} = e^{-ac}. \tag{19}$$

Equation (19) implies that

$$t = a^2 \frac{(\text{Var}_1 \tilde{P}_2)\text{Var } i}{(1+M)^2}$$

is determined solely by ac and is an increasing function of ac. The lower the cost of maintaining a market presence, the greater the number of market makers in equilibrium. That number would also be larger, of course, the smaller the risk-aversion parameter a for the market makers.

Var i is the average size of hedging demand (since hedging demand in its average size is $\text{E}|\,i\,|$ which is proportional to Var i when $\tilde{i}$ is normally distributed). $\text{Var}_1 \tilde{P}_2$ is the predictability of the price change. Hence, as either of these two variances rises the number of market makers rises.

C. *Some Empirical Implications of the Model*

The contribution of market makers shows up in the correlation between successive price changes. Since the model is only a three-period model with a single liquidity event at time 1, we define the correlation to be

$$q = \frac{\text{Cov}(P_2 - P_1,\ P_1 - \text{E}_0 P_1)}{\sqrt{\text{Var}(P_2 - P_1)\text{Var}(P_1 - \text{E}_0 P_1)}}.$$

Using (14), the fact that $P_2 = \text{E}_2 \tilde{P}_3$, $\text{E}_0 \tilde{P}_1 = \text{E}_0 \tilde{P}_2$ and $\text{E}_1 \tilde{P}_2 = \text{E}_1 \tilde{P}_3$ yields

$$P_2 - P_1 = P_2 - \text{E}_1 \tilde{P}_2 + \frac{i}{1+M}\, a\, \text{Var}_1(\tilde{P}_2) \tag{21}$$

$$P_1 - \text{E}_0 P_1 = \text{E}_1 \tilde{P}_2 - \text{E}_0 \tilde{P}_2 - \frac{i}{1+M}\, a\, \text{Var}_1(\tilde{P}_2). \tag{22}$$

To impart a timeless quality to the uncertainty, assume that one-step-ahead variances are the same at each date, i.e.,

$$s^2 = \text{Var}_1(\tilde{P}_2 - \text{E}_1 \tilde{P}_2) = \text{Var}(\text{E}_1 \tilde{P}_2 - \text{E}_0 \tilde{P}_2) = \text{Var}_1(\tilde{P}_2).$$

We can now restate q:

$$q = -\frac{t}{1+t}. \tag{23}$$

Thus, from (19) the correlation between successive price changes is negative and is determined solely by the cost of being a market maker c.

Note that the covariance between successive price changes is

$$\text{Cov}(\tilde{P}_2 - P_1,\ \tilde{P}_1 - \text{E}_0 \tilde{P}_1) = -\frac{a^2 \text{Var } i}{(1+M)^2}\, s^4 = -ts^2. \tag{24}$$

Hence, for a given c, since t is fixed, assets with more variability of expected price changes will have higher negative covariance.

Finally, consider the amount of immediacy provided in equilibrium. This can be measured by the amount of customer trade that is completed in period 1, x_1^{cd}, and the amount completed in period 2, $x_2^{cd} - x_1^{cd}$, which can be derived from (11)–(14):

$$x_1^{cd} = -\frac{M}{1+M}\, i \tag{25a}$$

$$x_2^{cd} - x_1^{cd} = -\frac{i}{1+M}. \tag{25b}$$

Since the total size of the trade desired is $-i$, the fraction completed in period 1 is determined by M. When M is very large the transaction is completed immediately and the market can be said to be liquid.

II. Extensions and Applications

Many readers will have been surprised to have come so far in a paper on market liquidity with no reference to the term "bid-ask spread." That term has indeed dominated academic discussions of transaction costs and market efficiency ever since the pioneering paper by Demsetz [3]; even before that, the term was the standard short-hand among practitioners for contrasting the cost of trading between markets and over time. For all its familiarity, however, and its rough common sense as a metric, we believe it does not fully capture the notion of market liquidity.

A. Limitations of the Bid-Ask Spread as a Measure of Liquidity

First (as Stoll [13] has emphasized), the bid-ask spread measures exactly the market maker's return for providing immediacy only in the special case in which the market maker simultaneously "crosses" (i.e., executes both sides of) the trade, one at the bid and the other at the ask. But in that case, of course, the spread could not also serve as a valid measure of the cost of supplying immediacy to each of its customers; it is simply a charge by the market maker for executing their orders, rather than for providing them liquidity services.

In the more typical case that our model was designed to portray, the orders do not arrive simultaneously but are randomly separated in time. If so, the price may change between the time at which the market maker buys and sells, and the market maker may earn much more or less than the spread quoted at the time of the first leg of the transaction. And, for the same reason, the currently quoted spread cannot serve any transactor as a precise measure of the cost of trading immediately rather than delaying the order, particularly when the order is a large one. Yet that cost, as we have emphasized, is the essence of market liquidity. A customer desiring to sell is likely to be more concerned with how the bid will change over time than with the size of the current bid-ask spread.

The benefit of immediacy to a customer is the shedding of the price risk associated with waiting. In most real-world exchanges this waiting can also be achieved by means of a "limit order" to sell, for example, at the current quoted bid price. Such a limit order will be executed if and when a buyer willing to pay this price appears and no other seller is offering to sell at a lower price. But that may never occur and the customer may have to revise the order and sell at a price lower than the bid price at the time the first limit order was sent in. Thus, if lucky, the limit-order customer gets a price higher than the bid, while if unlucky, a lower price. The customer's choice between limit orders and market orders is thus governed not by the bid-ask spread, but precisely by those considerations that our model suggests determine the supply and demand for immediacy, i.e., by the likelihood that a buyer will arrive who is willing to pay more than the current bid. (See Cohen, et al. [1] for an equilibrium analysis of bid-ask spreads that emphasizes the importance of jumps in the price away from the current quotes.)

Note also that a substantial volume of transactions occurs within the prevailing quoted bid-ask spread because the traders who commit to a bid (or ask) are giving the market an option. Some traders may decide not to commit to buying or selling at particular prices and thus the quoted bid may be lower than the actual bids that appear in response to a market sell order.

The more that market orders to buy and sell are separated in time, the greater the exposure of the market maker to the risks of adverse information trading. The bid-ask spread, in addition to the pure timing-option premium, will then contain still another component, which compensates the market makers on their informationless trades for their likely losses to the informed traders. This phenomenon, as noted earlier, has been much studied in the academic literature on market microstructure. (See, e.g., Glosten and Milgrom [5].) Much less attention, however, has been directed to the inverse problem of what is likely to happen to conventional quoted bid-ask spreads in highly active markets, like futures markets, where many separate buy and sell orders are entering the trading pit virtually simultaneously. Because the adverse-selection problem arises only when a market maker cannot hope to offset a position immediately, and because the costs of maintaining a market presence are mainly (and, in our model, entirely) fixed costs, it might seem that quoted bid-ask spreads and market-makers' profits from what amounts to crossing trades would be driven towards zero by the competitive entry of new market makers.[6] Where the fixed costs are large relative to the entry-inhibiting trading risks, a competitive market may not be viable because the market makers would have no way of recovering their fixed costs of maintaining a presence on the floor.[7] To keep markets viable, therefore, exchanges may limit the number of "seats" available to market makers (or designate a regulated specialist).

[6] Remember that, in our model, market makers take risky positions as well as match orders. Entry occurs to the point where the market makers earn a return on their risky positions plus any profits from simultaneous matching that just balances the trading risks and the fixed costs of maintaining a continuous presence.

[7] In terms of the notation in our model, the non-viability of a competitive equilibrium would occur when c becomes large relative to a.

Exchanges also typically define a minimum price-change unit (called a "tick") which, in highly active markets, serves also to set a minimum on both the quoted bid-ask spread and the profits a "scalper" makes from a quick turnaround. This somewhat subtle and frequently overlooked role of the minimum tick helps explain, among other things, the seeming paradox of finding many traders in an obviously highly competitive pit fighting (sometimes literally) to execute an order. This behavior suggests that the quoted bid-ask spread of one tick, and hence the profit from a quick turn on a standard-size trade is actually higher in an active market than it would be in the absence of the minimum-tick rule. Part of the art of managing a futures exchange is finding a minimum tick size for its contracts, high enough to sustain a viably competitive supply of floor traders, but not so high as to give rise to the problems of rationing and queue discipline so often encountered under price controls.[8]

B. Limitations of the "Liquidity Ratio" as a Measure of Market Liquidity

Another widely used empirical measure in inter-market comparisons of market liquidity is the "liquidity ratio," defined as the ratio of average dollar volume of trading to the average price change during some interval. (See, e.g., Dubofsky and Groth [4], Cooper, Groth and Avers [2], and Martin [10].) A high value for the ratio is taken to indicate that many shares were traded with little price change, and a low value is taken to suggest that a trader bringing a large block to market will induce a large adverse price change.

These measures, of course, tell us at best only about past average associations between price changes and volume. They do not answer the critical question of how the sudden arrival of a larger-than-average order would affect price. Nor do they distinguish adequately among the sources of price volatility. A particular market may display high price variability not because it is illiquid but because new fundamental information arrives frequently. High price volatility can occur without high volumes of trading; in fact, when the import of the news is unambiguous, there may be no trading at all.

The liquidity ratio, in sum, fails to capture what we have called the immediacy that the market's structure offers. At best, and with all due regard for the pitfalls of estimating simultaneous equations, it might hope to measure the average elasticity of the market's demand curve for transactions. What we need, however, is a measue of how well the market makers are providing customers with an effective substitute for the delays in a search for a more inclusive set of counterparties. Whether so complex a notion can ever be distilled down to a single scalar is still far from clear. Our equations (24) and (25) (a) and (b) with their focus on reversals offer some promising new leads (similar in spirit to those opened earlier by Roll [12]), which we hope to follow up in future empirical research.

The need for new ways of measuring and comparing the liquidity of different

[8] A closely related but somewhat different problem is faced by the designers of computerized, automatic execution systems like the much-publicized (but little used) INTEX exchange in Bermuda. Because the users can hit directly any bids or offers showing on the screen, no intermediary can hope to earn a living by "scalping" the bid-ask spread on quick trades. This keeps market makers, who might otherwise provide immediacy when orders do not match, from being able to recover their opportunity costs of maintaining a continuous presence in the market.

market structures takes on added urgency in the light of the dramatic stock market events of last October and especially of the many policy proposals for market reform that have surfaced in the wake of the crash and are now being actively debated in the press and in Congress. But even in the absence of numerical calibrations of liquidity we believe that the model of market liquidity presented in the previous section can offer a helpful perspective on the main events of those hectic days.

C. Market Liquidity and the Crash of October 1987

We hasten to add that our interpretation of the recent crash in terms of our model of market liquidity must not be taken as signifying our belief that the event was entirely, or even primarily, a matter of liquidity rather than of fundamentals. (See Miller, et al. [11] for a discussion of the events preceding and surrounding the crash.) Whatever the precipitating cause, a massive liquidity event, in our earlier sense of an imbalance in the demand for immediacy, clearly occurred at the opening of the markets on the 19th. Both the futures market and the cash spot market were hit simultaneously with a flood of sell orders of unprecedented size.

Each of the two markets responded immediately to the imbalances, but in ways appropriate to their characteristic and, as we had noted earlier, quite different structures. The rules of the NYSE permit—indeed, encourage—specialists to delay the opening of trading when the overnight accumulation of orders for a particular stock is too far out of balance to allow market clearing at a price near the previous close. The delayed opening gives the specialist time to search the floor and the upstairs block-trading desks for balancing orders on the other side. Under ordinary conditions, when most other stocks have opened and are trading normally, that search is completed successfully, and trading resumed (though, typically, with a somewhat larger than usual price gap) in a matter of a few minutes. At the opening of the 19th, however, the order imbalances were so widespread and so large that no immediate help from on or off the floor was available to the beleaguered specialists of many of the most heavily traded shares. An hour after the opening bell, more than a third of the stocks in the Dow-Jones Index (including such widely followed international companies as IBM, Sears and Exxon) had yet to start trading.

By contrast, the S&P 500 futures market at the Chicago Mercantile Exchange, like other futures markets, seeks to provide a setting in which prices can most speedily reflect the best current information. If the outcry at the opening call on a futures exchange shows the overnight accumulation of orders to be heavily unbalanced, then the price will jump directly to a level at which trading can immediately take place.[9] The previous closing price plays no explicit role in

[9] The Chicago Board of Options Exchange opens with an auctioneer establishing provisional opening prices for each traded option. But with so many separate maturities and striking prices involved, the process of finding simultaneous, viable trading ranges is far from easy when prices are moving rapidly. On the morning of the 19th, and again on the 20th, by the time the "rotation," as the opening process is called, had worked its way around to the last series, the earlier, tentatively established trading ranges had become hopelessly wide of the mark. The process had to be repeated, and on Tuesday trading did not in fact begin until far after the regular opening time.

setting the level or the path to reach it. This contrast in opening procedures between the futures and the stock markets is fully understandable in the light of our model. The high demand for immediacy by firms that use futures markets to hedge inventory risk causes those markets to be organized precisely to provide maximal immediacy of order execution. The costs of delayed execution being normally less for stock trading, the market makers there seek to provide more search service relative to immediacy than in the futures markets.

On Monday the 19th of October, opening prices in New York had to fall some ten percent below the Friday close—an enormous gap by past standards—before trading in all stocks could begin.[10] By 11:00 A.M. or so, New York time, however, all the major delayed-opening stocks had resumed trading, and the two markets were now virtually back in step. Although the price fall had been large, the two markets, from all outward appearances, appeared to have handled successfully the enormous imbalance of sell orders that had accumulated at the opening. But the capital resources of their regular market makers on or around the floor had by then been heavily committed. In Chicago, many of the smaller market makers had left the floor, either voluntarily or under pressure from their clearing firms. Those that remained were unwilling to take on large positions in such a volatile market except at price concessions far larger than normal. When a further wave of sell orders hit both markets somewhat after noon, New York time, there was less resistance from the market makers and the fabled "meltdown" was soon under way. Or, to use the less colorful language of our model, both markets had by then become highly illiquid and virtually incapable of supplying immediacy at the low cost their users in the past had come to expect.

That illiquidity was evidenced in the spot market by (1) the virtual impossibility of executing market sell orders at the bid quoted at the time of order entry, and (2) the delays in executing and confirming trades on Monday afternoon and again, after the opening on Tuesday.[11] On the futures exchange, order flows that might have moved the market by at most a tick or two in the week before, were moving the market by ten or twenty times that amount or more in the early afternoon of Tuesday, October 20. Despite the evident rise in the cost of immediacy to sellers, the inflow of sell orders continued, and perhaps even accelerated in what took on all the appearances of a classic, self-reinforcing panic. By early afternoon on Tuesday, trading had been suspended in many NYSE stocks and in the main options and futures markets. With virtually no market-making capacity remaining, the burden of equilibration had to be assumed by the search for buyers off the market, culminating in the cavalry-like ride to the rescue on Tuesday afternoon by large U.S. corporations instituting buy-back

[10] This difference in opening procedures in the two markets undoubtedly contributed to the widespread (but misleading) impression at the time that the futures market in Chicago, if not actually dragging down stock prices in New York, was at least signalling to an already panicky public that heavy new selling pressure was on its way to the market in New York.

[11] In the case of the NASDAQ bulletin board, market prices were sometimes changing at a faster rate than the quotes were being updated. When the best offer to sell is entered below the best bid to buy, a market is deemed crossed, and under the then-standing NASDAQ rules only the bid showed on the screen. No further transactions could be made until the obsolete bid was updated which often involved substantial delay.

programs of their own shares. At the same time, the Federal Reserve System was directly and indirectly encouraging banks to support dealer inventory positions. By the end of the day, these infusions of buying power had pushed prices nearly back to their levels before the Monday noon collapse and substantial market-making capacity was back in place.

Effective market-making capacity in the period immediately after the crash, however, as well as at several critical junctures during the crash, was reduced by restrictions imposed on "program trading" which cut the normal arbitrage linkage between the market makers in the spot and futures markets. Arbitragers, by taking offsetting positions in both markets close to simultaneously, can transmit some of the pressure of order imbalances from the market first impacted to the market makers in the other. Market makers' resources in both markets can thus be brought to bear on the initiating imbalance more effectively, much as they would be if the number of active market makers had been increased. Price concessions and hence the cost of transacting can be kept smaller in both markets, thanks to arbitrage program trading, than might otherwise be the case.[12] How ironic then to find arbitrage program trading still so often blamed for undermining investor confidence in the market.

[12] For a further discussion of arbitrage program trading and especially its interaction with portfolio insurance, see Grossman [8].

REFERENCES

1. K. Cohen, S. Mair, R. Schwartz, and D. Whitcomb. "Transaction Costs, Order Placement Strategy and the Existence of the Bid-Ask Spread." *Journal of Political Economy* 89 (April 1981).
2. K. Cooper, J. C. Groth, and W. E. Avers. "Liquidity, Exchange Listing, and Common Stock Performance." Working paper, Texas A&M University, August 1983.
3. H. Demsetz. "The Cost of Transacting." *Quarterly Journal of Economics* 82 (February 1968).
4. F. Dubofsky and J. Groth. "Exchange Listing and Liquidity." Department of Finance, Texas A&M University, mimeo, February 1984.
5. Lawrence R. Glosten and Paul R. Milgrom. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14 (March 1985), 71–100.
6. Sanford J. Grossman and Merton H. Miller. "Economic Costs and Benefits of the Proposed One-Minute Time Bracketing Regulation." *Journal of Futures Markets* 6 (Spring 1986), 141–66.
7. ———. "The Determinants of Market Liquidity." Manuscript, University of Chicago, July 1986.
8. Sanford J. Grossman. "An Analysis of the Implications for Stock and Futures Price Volatility of Program Trading and Dynamic Hedging Strategies." *Journal of Business* (1988), forthcoming.
9. T. Ho. "Dealer Market Structure: A Dynamic Competitive Model." New York University working paper, March 1984.
10. P. Martin. "Analysis of the Impact of Competitive Rates on the Liquidity of NYSE Stocks." Economic Staff Paper 75-3, Securities and Exchange Commission (July 1975).
11. Merton H. Miller, John D. Hawke Jr., Burton Malkiel, and Myron Scholes. *Preliminary Report of the Committee of Inquiry Appointed by the Chicago Mercantile Exchange to Examine the Events Surrounding October 19, 1987* (December 22, 1987), mimeo.
12. Richard Roll. "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market." *Journal of Finance* 39 (March 1984), 1127–39.
13. Hans R. Stoll. "Alternative Views of Market Making." In Y. Amihud, T. Ho, and R. Schwartz (eds.), *Market Making and the Changing Structure of the Securities Industry.* Lexington Books, 1985, 67–92.
14. David Whitcomb. "An International Comparison of Stock Exchange Trading Structures." In Y. Amihud, T. Ho, and R. Schwartz (eds.), *Market Making and the Changing Structure of the Securities Industry.* Lexington Books, 1985, 237–56.