Market Microstructure

Hans R. Stoll

Owen Graduate School of Management
Vanderbilt University
Nashville, TN 37203
Hans.Stoll@Owen.Vanderbilt.edu

Financial Markets Research Center
Working paper Nr. 01- 16

First draft: July 3, 2001
This version: August 15, 2002
Corrected May 19, 2003

Forthcoming *Handbook of the Economics of Finance*, edited by G.M. Constantinides, M. Harris, and R. Stulz, 2003, Elsevier Science B.V.

JEL classification: G20, G24, G28, G10, G14.

Key words: bid-ask spread; price impact, market design, dealer market, auction market, short-run price behavior, market fragmentation.

Market Microstructure

Hans R. Stoll

Abstract

Market microstructure deals with the purest form of financial intermediation -- the trading of a financial asset, such as a stock or a bond. In a trading market, assets are not transformed but are simply transferred from one investor to another. The field of market microstructure studies the cost of trading securities and the impact of trading costs on the short-run behavior of securities prices. Costs are reflected in the bid-ask spread (and related measures) and in commissions. The focus of this chapter is on the determinants of the spread rather than on commissions. After an introduction to markets, traders and the trading process, I review the theory of the bid-ask spread in section II and examine the implications of the spread for the short run behavior of prices in section III. In section IV, the empirical evidence on the magnitude and nature of trading costs is summarized, and inferences are drawn about the importance of various sources of the spread. Price impacts of trading from block trades, from herding or from other sources, are considered in section V. Issues in the design of a trading market, such as the functioning of call versus continuous markets and of dealer versus auction markets, are examined in section VI. Even casual observers of markets have undoubtedly noted the surprising pace at which new trading markets are being established even as others merge. Section VII briefly surveys recent developments in U.S securities markets and considers the forces leading to centralization of trading in a single market versus the forces leading to multiple markets. Most of this chapter deals with the microstructure of equities markets. In section VIII, the microstructure of other markets is considered. Section IX provides a brief discussion of the implications of microstructure for asset pricing. Section X concludes.

Market Microstructure

Hans R. Stoll

Market microstructure deals with the purest form of financial intermediation -- the trading of a financial asset, such as a stock or a bond. In a trading market, assets are not transformed (as they are, for example, by banks that transform deposits into loans) but are simply transferred from one investor to another. The financial intermediation service provided by a market, first described by Demsetz (1968) is immediacy. An investor who wishes to trade immediately – a demander of immediacy – does so by placing a market order to trade at the best available price – the bid price if selling or the ask price if buying. Bid and ask prices are established by suppliers of immediacy. Depending on the market design, suppliers of immediacy may be professional dealers that quote bid and ask prices or investors that place limit orders, or some combination.

Investors are involved in three different markets – the market for information, the market for securities and the market for transaction services. Market microstructure deals primarily with the market for transaction services and with the price of those services as reflected in the bid-ask spread and commissions. The market for securities deals with the determination of securities prices. The literature on asset pricing often assumes that markets operate without cost and without friction whereas the essence of market microstructure research is the analysis of trading costs and market frictions. The market for information deals with the supply and demand of information, including the incentives of securities analysts and the adequacy of information. This market, while conceptually separate, is closely linked to the market for transaction services since the difficulty and cost of a trade depends on the information possessed by the participants in the trade.

Elements in a market are the investors who are the ultimate demanders and suppliers of immediacy, the brokers and dealers who facilitate trading, and the market facility within which trading takes place. Investors include individual investors and institutional investors such as pension plans and mutual funds. Brokers are of two types: upstairs brokers, who deal with investors, and downstairs brokers, who help process transactions on a trading floor. Brokers are agents and are paid by a commission. Dealers

trade for their own accounts as principals and earn revenues from the difference between their buying and selling prices. Dealers are at the heart of most organized markets. The NYSE (New York Stock Exchange) specialist and the Nasdaq (National Association of Securities Dealers Automated Quotation) market makers are dealers who maintain liquidity by trading with brokers representing public customers. Bond markets and currency markets rely heavily on dealers to post quotes and maintain liquidity.

The basic function of a market – to bring buyers and sellers together -- has changed little over time, but the market facility within which trading takes place has been greatly influenced by technology. In 1792, when the New York Stock Exchange was founded by 24 brokers, the market facility was the buttonwood tree under which they stood. Today the market facility, be it the NYSE, Nasdaq or one of the new electronic markets, is a series of high-speed communications links and computers through which the large majority of trades are executed with little or no human intervention. Investors may enter orders on-line, have them routed automatically to a trading location and executed against standing orders entered earlier, and automatically sent for clearing and settlement. Technology is changing the relationship among investors, brokers and dealers and the facility through which they interact.

Traditional exchanges are membership organizations for the participating brokers and dealers. New markets are computer communications and trading systems that have no members and that are for-profit businesses, capable in principal of operating without brokers and dealers. Thus while the function of markets – to provide liquidity to investors – will become increasingly important as markets around the world develop, the exact way in which markets operate will undoubtedly change.

The field of market microstructure deals with the costs of providing transaction services and with the impact of such costs on the short run behavior of securities prices. Costs are reflected in the bid-ask spread (and related measures) and commissions. The focus of this chapter is on the determinants of the spread rather than on commissions. After an introduction to markets, traders and the trading process, I review the theory of the bid-ask spread in section II and examine the implications of the spread for the short run behavior of prices in section III. In section IV, the empirical evidence on the magnitude and nature of trading costs is summarized, and inferences are drawn about the

importance of various sources of the spread. Price impacts of trading from block trades, from herding or from other sources, are considered in section V. Issues in the design of a trading market, such as the functioning of call versus continuous markets and of dealer versus auction markets, are examined in section VI. Even casual observers of markets have undoubtedly noted the surprising pace at which new trading markets are being established even as others merge. Section VII briefly surveys recent developments in U.S securities markets and considers the forces leading to centralization of trading in a single market versus the forces leading to multiple markets. Most of this chapter deals with the microstructure of equities markets. In section VIII, the microstructure of other markets is considered. Section IX provides a brief discussion of the implications of microstructure for asset pricing. Section X concludes.[1]

I. **Markets, traders and the trading process**.

I.A. Types of markets.

It is useful to distinguish major types of market structures, although most real-world markets are a mixture of market types. An important distinction is between auction and dealer markets. A pure **auction** market is one in which investors (usually represented by a broker) trade directly with each other without the intervention of dealers. A **call auction** market takes place at specific times when the security is called for trading. In a call auction, investors place orders – prices and quantities – which are traded at a specific time according to specific rules, usually at a single market clearing price. For example, the NYSE opens with a kind of call auction market in which the clearing price is set to maximize the volume of trade at the opening.

While many markets, including the NYSE and the continental European markets, had their start as call auction markets, such markets have become continuous auction markets as volume has increased. In a **continuous auction** market, investors trade against resting orders placed earlier by other investors and against the "crowd" of floor brokers. Continuous auction markets have two-sides: Investors, who wish to sell, trade at the bid

[1] For other overviews of the field of market microstructure, see Madhavan (2000), the chapter in this volume by Easley and O'Hara, and O'Hara (1995).

price established by resting buy orders or at prices in the "crowd," and investors, who wish to buy, trade at the asking price established by resting sell orders or at prices in the "crowd." The NYSE is said to be a continuous auction market with a "crowd". Electronic markets are continuous auction markets without a "crowd."

A pure **dealer** market is one in which dealers post bids and offers at which public investors can trade. The investor cannot trade directly with another investor but must buy at the dealers ask and sell at the dealers bid. Bond markets and currency markets are dealer markets. The Nasdaq Stock Market started as a pure dealer market, although it now has many features of an auction market because investors can enter resting orders that are displayed to other investors.

Dealer markets are physically dispersed and trading is conducted by telephone and computer. By contrast, auction markets have typically convened at a particular location such as the floor of an exchange. With improvements in communications technology, the distinction between auction and dealer markets has lessened. Physical centralization of trading on an exchange floor is no longer necessary. The purest auction market is not the NYSE, but an electronic market (such as Island or the Paris Bourse) that takes place in a computer. The NYSE, in fact is a mixed auction/dealer market because the NYSE specialist trades for his own account to maintain liquidity in his assigned stocks. The Nasdaq Stock market is in fact also a mixed dealer/auction market because public orders are displayed and may be executed against incoming orders.

I.B. Types of orders

The two principal types of orders are a market order and a limit order. A **market order** directs the broker to trade immediately at the best price available. A **limit order** to buy sets a maximum price that will be paid, and a limit order to sell sets a minimum price that will be accepted. In a centralized continuous auction market, the best limit order to buy and the best limit order to sell (the top of the book) establish the market, and the quantities at those prices represent the depth of the market. Trading takes place as incoming market orders trade with the best posted limit orders. In traditional markets, dealers and brokers on the floor may intervene in this process. In electronic markets the process is fully automated.

In a pure dealer market, limit orders are not displayed but are held by the dealer to whom they are sent, and market orders trade at the dealers bid or ask, not with the limit orders. In some cases, such as Nasdaq before the reforms of the mid 1990s, a limit order to buy only executes if the dealer's ask falls to the level of the limit price. For example suppose the dealer's bid and ask are 20 to 20 ¼ , and suppose the dealer holds a limit order to buy at 20 1/8. Incoming sell market orders would trade at 20, the dealer bid, not at 20 1/8, the limit order. The limit order to buy would trade only when the ask price fell to 20 1/8. Nasdaq rules have been modified to require that the dealer trade customer limit orders at the same or better price before trading for his own account (Manning Rule), and to require the display of limit orders (the SEC's order handling rules of 1997).

Orders may also be distinguished by size. Small and medium orders usually follow the standard process for executing trades. Large orders, on the other hand, often require special handling. Large orders may be "worked" by a broker over the course of the day. The broker uses discretion when and how to trade segments of the order. Large orders may be traded in blocks. Block trades are often pre-negotiated "upstairs" by a broker who has identified both sides of the trade. The trade is brought to a trading floor, as required by exchange rules and executed at the pre-arranged prices. The exchange specifies the rules for executing resting limit orders.

I.C. Types of traders

Traders in markets may be classified in a variety of ways.

Active versus passive

Some traders are active (and normally employ market orders), while others are passive (and normally employ limit orders). Active traders demand immediacy and push prices in the direction of their trading, whereas passive traders supply immediacy and stabilize prices. Dealers are typically passive traders. Passive traders tend to earn profits from active traders.

Liquidity versus informed

Liquidity traders trade to smooth consumption or to adjust the risk-return profiles of their portfolios. They buy stocks if they have excess cash or have become more risk tolerant, and they sell stocks if they need cash or have become less risk tolerant. Informed

traders trade on private information about an asset's value. Liquidity traders tend to trade portfolios, whereas informed traders tend to trade the specific asset in which they have private information. Liquidity traders lose if they trade with informed traders. Consequently they seek to identify the counterparty. Informed traders, on the other hand, seek to hide their identity. Many models of market microstructure involve the interaction of informed and liquidity traders.

Individual versus institutional

Institutional investors – pension funds, mutual funds, foundations and endowments – are the dominant actors in stock and bond markets. They hold and manage the majority of assets and account for the bulk of share volume. They tend to trade in larger quantities and face special problems in minimizing trading costs and in benefiting from any private information. Individual investors trade in smaller amounts and account for the bulk of trades. The structure of markets must accommodate these very different players. Institutions may wish to cross a block of 100,000 shares into a market where the typical trade is for 3,000 shares. Markets must develop efficient ways to handle the large flow of relatively small orders while at the same time accommodating the needs of large investors to negotiate large transactions.

Public versus professional

Public traders trade by placing an order with a broker. Professional traders trade for their own accounts as market makers or floor traders and in that process provide liquidity. Computers and high speed communications technology have changed the relative position of public and professional traders. Public traders can often trade as quickly from upstairs terminals (supplied to them by brokers) as professional traders can trade from their terminals located in offices or on an exchange floor. Regulators have drawn a distinction between professional and public traders and have imposed obligations on professional traders. Market makers have an affirmative obligation to maintain fair and orderly markets, and they are obligated to post firm quotes. However, as the distinction between a day trader trading from an upstairs terminal and a floor trader becomes less clear, the appropriate regulatory policy becomes more difficult.

I.D. Rules of precedence

Markets specify the order in which resting limit orders and/or dealer quotes execute against incoming market orders. A typical rule is to give first priority to orders with the best price and secondary priority to the order posted first at a given price. Most markets adhere to price priority, but many modify secondary priority rules to accommodate large transactions. Suppose there are two resting orders at a bid price of $40. Order one is for 2,000 shares and has time priority over order two, which is for 10,000 shares. A market may choose to allow an incoming market order for 10,000 shares to trade with resting order two rather than break up the order into multiple trades. Even price priority is sometimes difficult to maintain, particularly when different markets are involved. Suppose the seller of the 10,000 shares can only find a buyer for the entire amount at $39.90, and trades at that price. Such a trade would "trade-through" the $40 price of order one for 2,000 shares. Within a given market, such trade-throughs are normally prohibited – the resting limit order at $40 must trade before the trade at $39.90. In a dealer market, like Nasdaq, where each dealer can be viewed as a separate market, a dealer may not trade through the price of any limit order he holds, but he may trade through the price of a limit order held by another dealer. When there are many competing markets each with its own rules of precedence, there is no requirement that rules of precedence apply across markets. Price priority will tend to rule because market orders will seek out the best price, but time priority at each price need not be satisfied across markets.

The working of rules of precedence is closely tied to the tick size, the minimum allowable price variation. As Harris (1991) first pointed out, time priority is meaningless if the tick size is very small. Suppose an investor places a limit order to buy 1000 shares at $40. If the tick size is $0.01, a dealer or another trader can step in front with a bid of 40.01 – a total cost of only $10. On the other hand, the limit order faces the danger of being "picked off" should new information warrant a lower price. If the tick size were $0.10, the cost of stepping in front of the investor's limit order would be greater ($100). The investor trades off the price of buying immediately at the current ask price, say $40.20, against giving up immediacy in the hope of getting a better price with the limit order at $40. By placing a limit order the investor supplies liquidity to the market. The

smaller tick size reduces the incentive to place limit orders and hence adversely affects liquidity.

Price matching and payment for order flow are other features of today's markets related to rules of precedence. Price matching occurs when market makers in a satellite market promise to match the best price in the central market for orders sent to them rather than to the central market. The retail broker usually decides which market maker receives the order flow. Not only is the broker not charged a fee, he typically receives a payment (of one to two cents a share) from the market maker. Price matching and payment for order flow are usually bilateral arrangements between a market making firm and a retail brokerage firm. Price matching violates time priority: When orders are sent to a price matching dealer, they are not sent to the market that first posted the best price. Consequently the incentive to post limit orders is reduced because the limit order may be stranded. Similarly, the incentive of dealers to post good quotes is eliminated if price matching is pervasive: A dealer who quotes a better price is unable to attract additional orders because orders are preferenced to other dealers who match the price.

I.E. The trading process

The elements of the trading process may be divided into four components – information, order routing, execution, and clearing. First, a market provides **information** about past prices and current quotes. Earlier in its history, the NYSE jealously guarded ownership of its prices, making data available only to its members or licensed recipients. But today transaction prices and quotes are disseminated in real-time over a consolidated trade system (CTS) and a consolidated quote system (CQS). Each exchange participating in these systems receives tape revenue for the prices and quotes it disseminates. The real-time dissemination of these prices makes all markets more transparent and allows investors to determine which markets have the best prices, thereby enhancing competition.

Second, a mechanism for **routing orders** is required. Today brokers take orders and route them to an exchange or other market center. For example, the bulk of orders sent to the NYSE are sent via DOT (Designated Turnaround System), an electronic system that sends an order directly to the specialist. Retail brokers establish procedures

for routing orders and may route orders in return for payments. Orders may not have the option of being routed to every trading center and may therefore have difficulty in trading at the best price. Central to discussions about a national market system, is the mechanism for routing orders among different market centers, and the rules, if any, that regulators should establish.

The third phase of the trading process is **execution**. In today's automated world this seems a simple matter of matching an incoming market order with a resting quote. However this step is surprisingly complex and contentious. Dealers are reluctant to execute orders automatically because they fear being "picked off" by speedy and informed traders, who have better information. Instead, they prefer to delay execution, if even for only 15 seconds, to determine if any information or additional trades arrive. Automated execution systems have been exploited by speedy customers to the disadvantage of dealers. Indeed, as trading becomes automated the distinction between dealers and customers decreases because customers can get nearly as close to "the action" as dealers.

A less controversial but no less important phase of the trading process is **clearing and settlement**. Clearing involves the comparison of transactions between buying and selling brokers. These comparisons are made daily. Settlement in U.S. equities markets takes place on day t+3, and is done electronically by book entry transfer of ownership of securities and cash payment of net amounts to the clearing entity.

II. Microstructure theory – determinants of the bid-ask spread.

Continuous markets are characterized by the bid and ask prices at which trades can take place. The bid-ask spread reflects the difference between what active buyers must pay and what active sellers receive. It is an indicator of the cost of trading and the illiquidity of a market. Alternatively, illiquidity could be measured by the time it takes optimally to trade a given quantity of an asset [Lippman and McCall (1986)]. The two approaches converge because the bid-ask spread can be viewed as the amount paid to someone else (i.e. the dealer) to take on the unwanted position and dispose of it optimally. Our focus is on the bid-ask spread. Bid-ask spreads vary widely. In inactive markets – for example, the real estate market – the spread can be wide. A house could be

offered at $500,000 with the highest bid at $450,000. On the other hand the spread for an actively traded stock is today often less than 10 cents per share. A central issue in the field of microstructure is what determines the bid-ask spread and its variation across securities.

Several factors determine the bid-ask spread in a security. First, suppliers of liquidity, such as the dealers who maintain continuity of markets, incur **order handling costs** for which they must be compensated. These costs include the costs of labor and capital needed to provide quote information, order routing, execution, and clearing. In a market without dealers, where limit orders make the spread, order handling costs are likely to be smaller than in a market where professional dealers earn a living. Second the spread may reflect **non competitive pricing**. For example, market makers may have agreements to raise spreads or may adopt rules, such as a minimum tick size, to increase spreads. Third, suppliers of immediacy, who buy at the bid or sell at the ask, assume **inventory risk** for which they must be compensated. Fourth, placing a bid or an ask grants an **option** to the rest of the market to trade on the basis of new information before the bid or ask can be changed to reflect the new information. Consequently the bid and ask must deviate from the consensus price to reflect the cost of such an option. A fifth factor has received the most attention in the microstructure literature; namely the effect of **asymmetric information**. If some investors are better informed than others, the person who places a firm quote (bid or ask) loses to investors with superior information.

The factors determining spreads are not mutually exclusive. All may be present at the same time. The three factors related to uncertainty – inventory risk, option effect and asymmetric information – may be distinguished as follows. The inventory effect arises because of possible adverse public information after the trade in which inventory is acquired. The expected value of such information is zero, but uncertainty imposes inventory risk for which suppliers of immediacy must be compensated. The option effect arises because of adverse public information before the trade and the inability to adjust the quote. The option effect really results from an inability to monitor and immediately change resting quotes. The adverse selection effect arises because of the presence of private information before the trade, which is revealed sometime after the trade. The information effect arises because some traders have superior information.

The sources of the bid-ask spread may also be compared in terms of the services provided and the resources used. One view of the spread is that it reflects the cost of the services provided by liquidity suppliers. Liquidity suppliers process orders, bear inventory risk, using up real resources. Another view of the spread is that it is compensation for losses to informed traders. This informational view of the spread implies that informed investors gain from uninformed, but it does not imply that any services are provided or that any real resources are being used.

Let us discuss in more detail the three factors that have received most attention in the microstructure literature – inventory risk, free trading option, and asymmetric information.

II.A. Inventory risk

Suppliers of immediacy that post bid and ask prices stand ready take on inventory and to assume the risk associated with holding inventory. If a dealer buys 5000 shares at the bid, she risks a drop in the price and a loss on the inventory position. An investor posting a limit order to sell 1000 shares at the ask faces the risk that the stock he is trying to sell with the limit order will fall in price before the limit order is executed. In order to take the risk associated with the limit order, the ask price must be above the bid price at which he could immediately sell by enough to offset the inventory risk. Inventory risk was first examined theoretically in Garman (1976), Stoll (1978a), Amihud and Mendelson (1980), Ho and Stoll (1981, 1983). This discussion follows Stoll (1978a).

To model the spread arising from inventory risk, consider the determination of a dealer's bid price. The bid price must be set at a discount below the consensus value of the stock to compensate for inventory risk. Let P be the consensus price, let P^b be the bid price, and let C be the dollar discount on a trade of Q dollars. The proportional discount of the bid price from the consensus stock price, P, is $\frac{P - P^b}{P} = \frac{C}{Q} \equiv c$. The problem is to derive C or equivalently, c. This can done by solving the dealer's portfolio problem. Let the terminal wealth of the dealer's optimal portfolio in the absence of making markets be $\tilde{W}$. The dealer's terminal wealth if he stands ready to buy Q dollars of stock at a discount of C dollars is $\tilde{W} + (1+\tilde{r})Q - (1+r_f)(Q-C)$, where $\tilde{r}$ is the return on the stock

purchased and r_f is the cost of borrowing the funds to buy the stock.[2] The minimum discount that the dealer would set is such that the expected utility of the optimal portfolio without buying the stock equals the expected utility of the portfolio with the unwanted inventory:

$$EU[\tilde{W}] = EU[\tilde{W} + (1+\tilde{r})Q - (1+r_f)(Q-C)]. \quad (1)$$

Applying a Taylor series expansion to both sides, taking expectations, assuming r_f is small enough to be ignored, and solving for $c=C/Q$, yields

$$c = \tfrac{1}{2}\frac{z}{W_0}s^2 Q \quad (2)$$

where z is the dealer's coefficient of relative risk aversion, W_0 is the dealer's initial wealth, σ^2 is the variance of return of the stock. The bid price for depth of Q dollars must be below the consensus stock value by the proportion c to compensate the dealer for his inventory costs. These costs arise because the dealer loses diversification and because he assumes a level of risk that is inconsistent with his preferences. The discount of the bid price is greater the greater dealer's risk aversion, the smaller his wealth, the greater the stock's return variance,[3] and the larger the quoted depth.

The proportional discount, c, is affected by the initial inventory of the dealer, which was assumed to be zero in the above derivation. If the dealer enters the period with inventory of I dollars in one or more stocks, the proportional discount for depth of Q can be shown to be

$$c = \frac{z}{W_0}s_{IQ}I + \tfrac{1}{2}\frac{z}{W_0}s^2 Q, \quad (3)$$

where s_{IQ} is the covariance between the return on the initial inventory and the return on the stock in which the dealer is bidding. If $I<0$ and $s_{IQ} > 0$, the dealer may be willing to pay a premium to buy shares because they hedge a short position in the initial inventory. On the other hand, the dealer's asking price will be correspondingly higher with an initial short position because the dealer will be reluctant to sell and add to the short position.

[2] Q is valued at the consensus price in the absence of a bid-ask spread. The loan is collateralized by the dealer's stock position.

[3] The variance, not the beta, is relevant because the inventory position is not diversified.

The relation between the bid price and consensus price for depth of Q and initial inventory of I is given by

$$\frac{P - P^b}{P} = \frac{z}{W_0} s_{IQ} I + \tfrac{1}{2} \frac{z}{W_0} s^2 Q , \qquad (4)$$

and the relation between the ask price and the consensus price for depth of Q and initial inventory of I is given by

$$\frac{P^a - P}{P} = -\frac{z}{W_0} s_{IQ} I + \tfrac{1}{2} \frac{z}{W_0} s^2 Q . \qquad (5)$$

Note that the inventory term enters with a negative sign in the ask equation since a positive value of *I* will lower the price a dealer will ask. (Q is an absolute dollar amount long or short). The proportional bid-ask spread if inventory costs were the only source of the spread is then given by summing (4) and (5):

$$\frac{P^a - P^b}{P} = 2c = \frac{z}{W_0} s^2 Q , \qquad (6)$$

Note that the initial inventory does not appear in the spread expression. Initial inventory affects the placement of the bid and ask but not the difference between the two. The implication for the dynamics of the quotes is that after a sale at the bid, both the bid and the ask price are lowered. The bid is lowered to discourage additional sales to the dealer, and the ask is lowered to encourage purchases from the dealer. Correspondingly, after a purchase at the ask, both bid and ask prices are raised.

The inventory model can be extended to account for multiple stocks, multiple dealers, and multiple time periods, without altering the essential features underlying the inventory approach.

II.B. Free trading option

A dealer or limit order placing a bid offers a free put option to the market, a fact first noted by Copeland and Galai (1983). For example, suppose an investor places a limit order to buy 5000 shares at a price of $40 when the last trade was at $40.25. The limit order gives the rest of the market a put option to sell 5000 shares at an exercise price of $40, which will be exercised if new information justifies a price less than $40. Similarly a limit order to sell at $40.50 offers a call option to the rest of the market, which will be

exercised if new information justifies a price greater than $40.50. A dealer who places a bid at $40 and an ask at $40.50 is writing a strangle. The value of such options depends on the stock's variability and the maturity of the option. A limit order that is monitored infrequently has greater maturity than a dealer quote that is monitored continuously and is quickly adjusted.

The Black Scholes model can provide the value of the free trading option. Suppose the limit order to buy at $40 will not be reviewed for an hour, and suppose the one-hour standard deviation of return is 0.033%. (an annualized value of about 200%). The stock price is $40.25 and the exercise price of the put option is $40. The Black Scholes value of a put option maturing in one hour with a one hour standard deviation of 0.033% is $0.23, approximately the discount of the bid from the quote midpoint. Investors who place limit orders expect to trade at favorable prices that offset the losses when their options end up in the money. The option is free to the person exercising it. The option premium, which is the discount of the bid from the stock's consensus value, is paid by traders who sell at the bid in the absence of new information.

II.C. Adverse selection.

Informed investors will sell at the bid if they have information justifying a lower price. They will buy at the ask if they have information justifying a higher price. In an anonymous market, dealers and limit orders must lose to informed traders, for the informed traders are not identified. If this adverse selection problem is too great the market will fail. As Bagehot (1971) first noted, the losses to informed traders must be offset by profits from uninformed traders if dealers are to stay in business and if limit orders are to continue to be posted. Glosten and Milgrom (1985) model the spread in an asymmetric information world. Important theoretical papers building on the adverse selection sources of the spread include Kyle (1985), Easley and O'Hara (1987), and Admati and Pfleiderer (1988).

The determination of the bid-ask spread in the Glosten/Milgrom world can illustrated in the following simple manner. Assume an asset can take on two possible values – a high value, v^H, and a low value, v^L -- with equal probability. Informed investors, who know the correct value, are present with probability π. Assuming risk

neutrality, uninformed investors value the asset at $\overline{v} = (v^H + v^L)/2$. The ask price, A, is then the expected value of the asset conditional on a trade at the ask price:

$$A = v^H p + \overline{v}(1-p). \tag{7}$$

The bid price is

$$B = v^L p + \overline{v}(1-p). \tag{8}$$

Since informed investors trade at the ask (bid) only if they believe the asset value is v^H (v^L), the ask price exceeds the bid price. The bid-ask spread,

$$A - B = p(v^H - v^L), \tag{9}$$

depends on the probability of encountering an informed trader and on the degree of asset value uncertainty. Glosten and Milgrom go on to show that prices evolve through time as a martingale, reflecting at each trade the information conveyed by that trade.

III. Short-run price behavior and market microstructure

Market microstructure is the study of market friction. In cross section, assets with greater friction have larger spreads. Friction also affects the short-term time series behavior of asset prices. Assets with greater friction tend to have greater short run variability of prices. Garman (1976) first modeled microstructure dynamics under the assumption of Poisson arrival of traders. Many papers have modeled the time series behavior of prices and quotes, including Roll (1984), Hasbrouck (1988, 1991), Huang and Stoll (1994, 1997), Madhavan, Richardson and Roomans (1997).

The evolution of prices through time provides insight as to the sources of trading friction – whether order processing costs, inventory effects, information effects, or monopoly rents.

- If order processing costs were the sole source of the bid-ask spread, transaction prices would simply tend to "bounce" between bid and ask prices. After a trade at the bid, the next price change would be zero or the spread, *S*. After a trade at the ask, the next price change would be zero or -*S*. Roll (1984) shows that the effect is to induce negative serial correlation in price changes.
- If asymmetric information were the sole source of the spread, transaction prices would reflect the information conveyed by transactions. Sales at the bid would cause

a permanent fall in bid and ask prices to reflect the information conveyed by a sale. Conversely purchases at the ask would cause a permanent increase in bid and ask prices to reflect the information conveyed by a purchase. Given the random arrival of traders, price changes and quote changes would be random and unpredictable.

- If inventory costs were the source of the spread, quotes would adjust to induce inventory equilibrating trades. After a sale at the bid, bid and ask prices would fall, not to reflect information as in the asymmetric information case, but to discourage additional sales and to encourage purchases. Correspondingly, after a purchase at the ask, bid and ask prices would rise to discourage additional purchases and to encourage sales. Over time, quotes would return to normal. Trade prices and quotes would exhibit negative serial correlation.

In this section, a model for examining short run behavior of prices is first presented. The model is then used to analyze the realized spread (what a supplier of immediacy earns) and the serial covariance of price changes. The realized spread and the serial covariance of price changes provide insight into the sources of the quoted spread.

III.A. A model of short term price behavior

The short run evolution of prices can be more formally stated. Let the change in the quote midpoint be given as

$$M_t - M_{t-1} = l\,\frac{S}{2}Q_{t-1} + e_t\ , \qquad (10)$$

where

M_t = quote midpoint immediately after the trade at time t-1.

Q_t = trade indicator for the trade at time t. Equals 1 if a purchase at the ask and equals -1 if a sale at the bid.

S = dollar bid-ask spread

λ = fraction of the half-spread by which quotes respond to a trade at t. The response reflects inventory and asymmetric information factors.

ε = serially uncorrelated public information shock.

The quote midpoint changes either because there is new public information, ε, or because the last trade, Q_{t-1} induces a change in quotes. A change in the quotes is induced because the trade conveys information and because it distorts inventory.

The trade at price P_t takes place either at the ask (half-spread above the midpoint) or at the bid (half-spread below the midpoint): [4]

$$P_t = M_t + \frac{S}{2} Q_t + h_t \ , \tag{11}$$

where

P_t = trade price at time t.

η_t = error term reflecting the deviation of the constant half-spread from the observed half-spread, $P_t - M_t$, and reflecting price discreteness.

Combining (10) and (11) gives

$$\Delta P_t = \frac{S}{2}(Q_t - Q_{t-1}) + l\,\frac{S}{2} Q_{t-1} + e_t \ , \tag{12}$$

where $e_t = e_t + Dh_t$.

III.B. The realized spread

What can a supplier of immediacy expect to realize by buying at the bid and selling his position at a later price (or by selling at the ask and buying to cover the short position at a later price)? The realized half-spread is the price change conditional on a purchase at the bid (or the negative of the price change conditional on a sale at the ask). Since quotes change as a result of trades, the amount earned is less than would be implied if quotes did not change. The difference between the realized and quoted spreads provides evidence about the sources of the spread.

In terms of the model (12), the expected realized half-spread conditional on a purchase at the bid $(Q_{t-1}=-1)$ is

[4] It would be a simple matter to model the fact that some trades take place inside the quotes. For example one could assume that trades are at the quotes with probability f and at the midpoint with probability (1-f). Then $P_t = M_t + f\,\frac{S}{2} Q_t + h_t$, Madhavan, Richardson, Roomans (1997), for example, make such an adjustment.

$$E\left[\Delta P_t \middle| Q_{t-1} = -1\right] = \frac{S}{2}(EQ_t + 1) + l\,\frac{S}{2}(-1), \qquad (13)$$

The expected realized half-spread depends on the expected sign of the next trade, EQ_t, and on λ. Let π be the probability of a reversal – a trade at the ask after a trade at the bid or a trade at the bid after a trade at the ask. Then, conditional on a trade at the bid, $E(Q_t) = p(1) + (1-p)(-1)$. If purchases and sales are equally likely, *EQ = 0.0,* (the liquidating transaction will be at midpoint on average). The value of λ depends on the presence of asymmetric information and/or inventory effects. The value of λ associated with alternative sources of the spread and the resulting values of *EQ* and of the realized spread are given in the following table:

Source of the spread	λ	E(Q)	Realized half-spread
Order processing	0	0	S/2
Asymmetric information	1	0	0
Inventory	1	2π-1	$(2\pi$-1)S/2

In an order processing world, *l = 0* because quotes are assumed not to adjust to trades, and *EQ = 0.0* because purchases and sales are assumed to arrive with equal probability. The implied realized half-spread is *S/2*, that is, the supplier of immediacy earns half the quoted spread. He would earn the spread on a roundtrip trade – buy at the bid and sell at the ask. These earnings defray the order processing costs of providing immediacy.

In an asymmetric information world, quotes adjust to reflect the information in the trade. If adverse information is the sole source of the spread, *l = 1.* A trade at the bid conveys adverse information with value S/2, causing quotes to decline by S/2. Since quotes reflect all current information, buys and sells continue to be equally likely so that *EQ = 0.0* at the new quotes. The resulting realized half-spread of zero reflects the fact that, in an asymmetric information world, real resources are not used up to supply immediacy and no earnings result. The spread is simply an amount needed to protect suppliers of immediacy from losses to informed traders.

In an inventory world, quotes also respond to a trade but not because the trade conveys information but because the trade unbalances the inventory of liquidity suppliers. If inventory is the sole source of the spread, $l = 1$. A trade at the bid causes quotes to decline by S/2. Since the fundamental value of the stock has not declined (as is the case in the asymmetric information case), the lower bid price makes it more costly to sell, and the lower ask price makes less expensive to buy. As a result, subsequent purchases and sales will not be equally likely. After a trade at the bid, a trade at the ask occurs with probability greater than 0.5, while a trade at the bid occurs with probability less than 0.5. For example if $p = 0.7$, $E(Q) = 0.4$, and the realized half-spread would be 0.4S/2. Given enough trades, quotes would return to their initial level, and the half-spread would be earned, but one is unlikely to observe a complete reversal in one trade.

A direct implication of the inventory world is that quote changes are negatively serially correlated, something that is not the case in the order processing world (where successive price changes, but not quote changes, are negatively correlated) or in the asymmetric information world (where neither price changes nor quote changes are serially correlated). The negative serial correlation in quotes tends to be long lived and the mean reversion of inventories tends to be slow, which makes inventory effects difficult to observe. [5] The serial covariance of price changes is examined in greater detail in the next section.

The above discussion has described polar cases. In fact, the sources of the quoted spread are likely to include order processing, asymmetric information, inventory, as well as market power and option effects. The relative importance of asymmetric effects and other effects can be inferred empirically by comparing the quoted half-spread and the realized half-spread. For example if the quoted half-spread were 10 cents, and suppliers of immediacy realized an average of 6 cents by buying at the bid (or selling at the ask) and liquidating their position at a later time, one would infer that the asymmetric portion of the half-spread is 4 cents and the other portions are 6 cents.

III.C. Serial covariance of price changes

[5] Madhavan and Smidt (1991, 1993) find that inventories are long lived. Hansch, Naik, Viswanathan (1998) find direct evidence of inventory effects in the London market.

Another approach to understanding the implications of market microstructure for price dynamics and the sources of the spread is to calculate the serial covariance of transaction price changes. This can be done by calculating the serial covariance of both sides of (12) under alternative assumption about λ. Consider first the order processing world, where $\lambda = 0$. Assuming in addition that markets are informationally efficient and that the error term is serially uncorrelated and uncorrelated with trades, implies that

$$\text{cov}(\Delta P_t, \Delta P_{t-1}) = \frac{S^2}{4}\text{cov}(\Delta Q_t, \Delta Q_{t-1}) = \frac{S^2}{4}(-4p^2)\,. \tag{14}$$

Assuming that the probabilities of purchases and sales are equal at π=0.5, the serial covariance of price changes is

$$\text{cov}(\Delta P_t, \Delta P_{t-1}) = -\frac{S^2}{4}\quad, \tag{15}$$

a result first derived by Roll (1984). For example, if *S = $0.20*, the serial covariance is –0.01. Roll pointed out that one could infer the spread from transaction prices as

$$S = 2\sqrt{-\text{cov}(\Delta P_t, \Delta P_{t-1})}\,. \tag{16}$$

Consider next the pure asymmetric information world or the pure inventory world, where $\lambda = 1$. In either of these cases,

$$\text{cov}(\Delta P_t, \Delta P_{t-1}) = \frac{S^2}{4}\text{cov}(Q_t, Q_{t-1}) = \frac{S^2}{4}(1-2p)\,.^{6} \tag{17}$$

In an asymmetric information world, since quotes are "regret free," they induce no serial dependence in trades and $\pi = 0.5$. In that case *(1-2p)= 0.0*, and $\text{cov}(\Delta P_t, \Delta P_{t-1}) = 0.0$.

In a pure inventory world, quote changes induce negative serial dependence in trading, that is to say π>0.5 (but is less than 1). The serial covariance in that case is

$$\text{cov}(\Delta P_t, \Delta P_{t-1}) = \frac{S^2}{4}(1-2p)\quad, \text{where } 0.5<p<1.0. \tag{18}$$

The serial covariance is negative but not as negative as in the pure order processing world in which *p =0.5*. The serial covariance is attenuated because quotes respond to trades. For example, if *S = 0.20, p = 0.7,* the serial covariance is –0.004.

---

[6] Note that the serial covariance in trade direction is $\text{cov}(Q_t, Q_{t-1}) = (1-2p)$ whereas the serial covariance in trade direction changes is $\text{cov}(\Delta Q_t, \Delta Q_{t-1}) = -4p^2$.

If the serial covariance is calculated from actual transaction prices and the Roll transformation applied, the inferred spread is typically less than the quoted spread. This happens for several reasons. First, as noted above, the response of quotes to trades because of information or inventory effects attenuates the bid-ask bounce. The serial covariance is less negative the more important the asymmetric information component of the spread. Second, the negative serial correlation in trades implied by microstructure theory comes from the supply side. However, investors' trading may be positively correlated. For example momentum trading implies $\text{cov}(\Delta P_t, \Delta P_{t-1}) > 0.0$. Positive demand side serial correlation may obscure or lessen negative serial correlation due to microstructure effects. Third, trade reporting procedures and price discreteness can obscure negative serial covariance implied by microstructure factors. For example, an investor's order may not be accomplished in a single trade but may be split into several trades all of the same sign. Breaking up an order in this way induces runs in the direction of trade and makes trade reversals less likely to be observed. Price discreteness can obscure price changes that might otherwise be observed and therefore can obscure serial correlation of price changes.

IV. Evidence on the bid-ask spread and its sources

IV.A. The spread and its components

Evidence on spreads for a sample of 1706 NYSE stocks in the three months ending in February 1998 is contained in Table 1. The **quoted half-spread** ranges from 8.28 cents per share for small, low priced stocks to 6.49 cents per share for large, high priced stocks, with an overall average of 7.87 cents per share. The higher spreads for small low priced stocks reflect the lesser liquidity of these stocks.

Row 2 of the table presents estimates of the **effective half-spread**. The effective spread is defined as $\left|P_t - M_t\right|$, the absolute difference between the trade price and the quote midpoint.[7] If the trade is at the bid or ask, the effective spread equals the quoted spread. However, because it is often possible for an incoming market order to better the

[7] This definition poses a number of empirical problems. First, to classify a trade, one must associate the trade price with the correct quotes, which can be problematic if there are differential reporting delays. Second, one must assume that trades above the midpoint are purchases and trades below the midpoint are sales. Lee and Ready (1991) analyze these questions.

quoted price, ("price improvement"), the effective spread may be less than the quoted spread. The process of achieving price improvement is for the dealer to guarantee the current price and seek to better it. Lee (1993) provides evidence on price improvement across different markets. Ready (1999) notes that the dealer has a very short term option, which is to step ahead of the resting order by bettering the price or to let the incoming market order trade against the resting order. Price improvement can adversely affect resting orders since dealers will likely step ahead if the incoming order is judged to be uninformed and will not step ahead if the incoming order is judged to be informed. The effective half-spread is below the quoted spread in each size category. It averages 5.58 cents over all NYSE stocks.

Both the quoted and effective spreads are measures of total execution cost, inclusive of real costs and of wealth transfers due to asymmetric information. A measure of real cost is the realized spread. Empirically, the realized spread may be estimated simply by calculating the average price change after a trade at the bid or the negative of the average price change after a trade at the ask. The price change is taken from the initial trade price to a subsequent price, where the subsequent price may be the quote midpoint or the trade price of a later trade. Huang and Stoll (1996) calculate realized spreads over 5 and 30 minute intervals. An alternative empirical estimate of the realized spread is to calculate half the average difference between trades at the ask and trades at the bid -- what Stoll (2000) has called the traded spread.

The relation between the average realized and traded half-spreads in a given day is as follows: The average **realized half-spread** for *m* trades taking place at bid prices is

$$\frac{1}{m}\sum_{T=1}^{m}(M_{T+1}-P_T^B), \tag{19}$$

where M_{T+1} is the quote midpoint at which the trade at time *T* is assumed to be liquidated and P_T^B is the bid price at which the trade at time *T* was initiated. The average realized spread for *n* trades taking place at ask prices is

$$-\frac{1}{n}\sum_{t=1}^{n}(M_{t+1}-P_t^A). \tag{20}$$

Note that the time subscripts (*t and T*) are different to reflect the fact that a trade at the bid and at the ask do not take place at exactly the same time. After each trade, the quotes

adjust to reflect the information in the trade and the inventory effects of the trade. Summing (19) and (20) gives

$$(\frac{1}{m}\sum_{T=1}^{m} M_{T+1} - \frac{1}{n}\sum_{t=1}^{n} M_{t+1}) + (\frac{1}{n}\sum_{t=1}^{n} P_t^A - \frac{1}{m}\sum_{T=1}^{m} P_T^B) \tag{21}$$

The **traded spread** is defined as

$$(\frac{1}{n}\sum_{t=1}^{m} P_t^A - \frac{1}{m}\sum_{T=1}^{m} P_T^B), \tag{22}$$

which is the same as (21) under the assumption that the midpoint at which trades are liquidated is the same for trades at the bid and trades at the ask. The traded spread is the average earnings of a supplier of immediacy who buys at the bid and sells at the ask. It is less than the quoted spread because prices tend to move against the supplier of liquidity after each trade.

The traded half-spread data in row 3 of Table 1 are based on weighted averages of trade prices where the weights are the volume at each price. As expected, the traded half-spread is less than the quoted half-spread, reflecting the fact that suppliers of immediacy earn less than the quoted spread primarily because they lose to informed traders. Over all NYSE stocks, the traded half-spread is 3.74 cents, which implies that losses to informed traders average 5.58 – 3.74 = 1.84 cents per share. Per share losses to informed traders are less in large stocks than in small stocks, reflecting the fact that there are many more shares traded in the large stocks.

The final measure summarized in Table 1 is the **Roll implied spread**, which is based on the serial covariance of price changes in each stock as given by (16). Like the traded spread, the Roll spread is less than the quoted or effective spread, reflecting the fact that asymmetric information lowers the earnings of suppliers of immediacy relative to the quoted or effective spread.

The comparison of the quoted and effective spreads with the realized spread, as represented by the traded spread or Roll spread, in Table 1 provides clear empirical support for the fact that a significant portion of the spread reflects the real costs of providing immediacy and a portion reflects the losses to informed trading. However the exact composition, and in particular the importance of inventory and asymmetric information effects is uncertain.

A number of authors have analyzed the components of the spread in greater detail and more formally than is possible with the simple comparisons in Table 1. Relevant studies include Glosten and Harris (1988), Stoll (1989), Choi, Salandro, Shastri (1988), George, Kaul and Nimalendran (1991), Lin, Sanger, Booth (1995), Huang and Stoll (1997).

IV.B. Cross section evidence.

Whatever the exact sources of the bid-ask spread, research has clearly established that the cross-section variation in spreads can be explained by economic variables. Indeed the relation between the spread of a security and trading characteristics of that security is one of the strongest and most robust relations in finance. The relation has been examined by Demsetz (1968), Stoll (1978b), Benston and Hagerman (1974), Branch and Freed (1977), Tinic (1972), Tinic and West (1974) and many other, more recent papers. Since most of the early empirical work preceded the articulation of the asymmetric information theories of the spread, the explanatory variables were based on inventory and order processing reasons, but in most cases asymmetric information factors can be represented by the same empirical proxies. Important variables include an activity variable like volume of trading, a risk variable like the stock's return variance, variables for company characteristics such as size and stock price that proxy for other aspects of risk, and perhaps other variables such as a variable for trading pressure, and a variable for price discreteness.

Results for the following cross section relation (taken from Stoll (2000) are in Table 2:

$$S/P = a_0 + a_1 \log V + a_2 s^2 + a_3 \log MV + a_4 \log P + a_5 \log N + a_6 Avg|I| + e \qquad (23)$$

The data are averages of daily data for 1706 NYSE/AMSE stocks for the 61 trading days ending February 28, 1998. The variables are as follows: S is the stock's average quoted half-spread defined as ½(ask price-bid price), P is the stock's average closing price, V is average daily dollar volume, s^2 is the daily return variance for the prior year, MV is the stock's market value at the end of November 1997, N is the average number of trades per day, I is the average daily percentage imbalance between volume at the ask and volume at

the bid, and *e* is the error term.[8] Over 79% of the cross section variation in proportional spreads is explained by stock characteristics. The key results are well known: spreads are lower for stocks with the greater volume, with lower return volatility, with higher price, and with smaller trading imbalances. The positive coefficient on the number of trades is somewhat surprising.

V. Price effects of trading

V.A. Block trading

Models of the bid-ask spread derive the prices at which suppliers of immediacy will buy (at the bid) or sell (at the ask) specified quantities (depth). Orders are assumed to be of a size less than or equal to the posted depth. Orders arrive and are executed at posted quotes, and quotes adjust to reflect information and inventory effects.

Institutional investors, such as mutual funds and pension funds, often must trade quantities that exceed the quoted depth. They are concerned about a price impact over and above that in the spread. An institution interested in selling 50,000 shares of a 40 dollar stock cannot simply place a market order. It has two options. First it can pre-negotiate the sale of the entire block in an upstairs market that is facilitated by major broker dealer firms. Second, it can ask a broker to "work" the order by trading portions of it throughout the day so as to minimize the price impact.

Block trades have been analyzed in a number of papers, including Scholes (1972), Kraus and Stoll (1972b), Holthausen et al (1987) and others. Markets regulate the interaction of block trades and ongoing trades. Suppose the current price of a stock is 40, and a block sale is negotiated upstairs at a price of 38. In the NYSE, the trade must be brought to the floor, where resting limit orders and floor brokers wishing to buy at 38 or more must be satisfied. Further the block trade must be reported publicly. By contrast, the London Stock Exchange has allowed reporting of the trade to be delayed up to 90 minutes in order to give broker dealers who acquire shares time to dispose of their shares. An alternative to crossing the block at 38 while the last trade took place at 40 is for the

[8] The above relation is only one of several possible formulations. For example, one could take the dollar spread as the dependent variable. Similarly, the independent variables can be expressed in alternative ways. The fundamental variables -- share volume, return variance, price, number of trades and market value -- almost always are strongly significant in each formulation.

broker to trade portions of the block at prices between 40 and 38 until the market price equals the pre-negotiated block price, and then trade the remaining block. The risk of pre-trading portions of the block in this manner is that other traders will become aware of the block and will sell in anticipation, perhaps driving the price down and forcing a lower block price.

The empirical evidence indicates that price impacts of block trading are quite mild. In part this reflects the ability of the broker to pre-trade and minimize the impact of the block. Kraus and Stoll (1972b) find a temporary price impact of 0.70% of the stock price for blocks that are sold and no temporary price impact for blocks that are purchased. The temporary price impact is akin to the bid-ask bounce of ordinary trades. The fact that prices do not bounce back after a block purchase implies that the price increase accompanying such blocks reflects new information. The asymmetry in price impacts between sale and purchase blocks is found in all block studies.

Since block trading is only one technique available to institutions, a natural issue is the overall trading costs of institutional investors. What are the impacts of institutional trading as seen from the perspective of institutions? A number of studies have gained access to institutional trading records in order to answer this question. These include Chan and Lakonishok (1993) and Keim and Madhavan (1997). An interesting feature of institutional trading data is that it is virtually impossible to connect institutional trade records to trades as reported over the tape. This is because institutions receive reports as to the average price of their trades in each stock on each day without a detailed breakdown as to the individual trades. Chan and Lakonishok report that buy programs have a price impact of 0.34 percent whereas sell programs have a price impact of only –0.04 percent.

V.B. Herding

Studies of individual institutions' trading do not assess the price impact of aggregate selling or buying pressure by several institutions. It is frequently said that institutions "herd" because they listen to the same analysts and go to the same clubs. In the first study of herding, Kraus and Stoll (1972a), using data collected as part of the

Institutional Investor Study[9], were able to construct monthly trading imbalances for the largest institutional investors in over 400 different stocks. They examine the tendency of institutions to trade in parallel and conclude that parallel trading does not occur more frequently than would be expected by chance. When parallel trading does occur, even though it be by chance, temporary price effects are observed. More recently, Lakonishok, Shleifer and Vishny (1992) analyze herding by pension funds. Wermers (1999) finds that mutual fund herding is related to past, contemporaneous and future returns. Mutual fund buying is more likely when past returns were positive, has a strong contemporaneous positive price effect, and tends to precede future positive returns.

An alternative approach to assessing the price impact of trading imbalances is to infer the imbalance from trade data. For a given day t, sell volume, S_t, is the number of shares traded below the quote midpoint, and buy volume, B_t, is the number of shares traded above the quote midpoint. The proportional imbalance on day t is $I_t = \frac{B_t - S_t}{B_t + S_t}$.

One approach to assessing the imbalance in a given stock is to estimate the following regression:

$$\Delta P_t = l_0 + l I_t + l_2 I_{t-1} + e_t \qquad (24)$$

where DP_t = stock's quote midpoint change (net of market) on day t. Use of the midpoint abstracts from the bid-ask bounce. The coefficient, l, measures the sensitivity of the quote change over a day to the daily imbalance. The coefficient is in the spirit of Kyle (1985). Insofar as the quote change is permanent, l measures the information content of the day's imbalance. If prices bounce back the next day, one would conclude that the price impact reflects real factors. Stoll (2000) estimates the above regression for 1706 NYSE stocks. Each stocks has 61 days of data. The value of λ is positive and highly significant, indicating that trading pressure affects prices. Easley, Kiefer, O'Hara and Paperman (1996) use data on trading pressures to infer the probability that an information event has occurred. In their model, an excess of sellers over buyers increases the probability that negative private information exists.

[9] See U.S. SEC, *Institutional Investor Study* (1971). Unlike later studies which rely on end of quarter holdings to infer purchases and sales, the data in the Institutional Investor Study are actual monthly purchases and sales provided by all major institutional investors over a period of 21 months.

V.C. Other studies of the effects of trading

A number of other studies have examined the relation between trading and the pattern of prices over time. French and Roll (1986) find that the variance of overnight returns (close to open) is only 1/5 the variance of daytime returns (open-to-close). While a large portion of the difference is due to the fact that news is not released during the night, they conclude that some of this difference is due to the fact that trading when the market is open causes volatility. Wood, McInish and Ord (1985) find that spreads are greatest in the morning, lowest at midday and increase somewhat at day-end, consistent with the fact that volatility is greatest around the opening. Harris (1989) and Madhavan, Richardson, Roomans (1997) have investigated the pattern of price behavior over the day, and Harris (1986) has investigated the pattern over days of the week. Over all, research on the time series pattern of spreads and volatility suggests that trading affects prices.

VI. Market design

Any securities market, be it a traditional membership organization like the NYSE or a new for-profit electronic market, must make some very practical decisions about how trading should be organized. Should the market be a call market or a continuous market? If continuous, how should the market open, and under what circumstances, if any, should trading be halted? Should the market be an order driven auction market that relies on limit orders to provide immediacy or should it be a quote driven dealer market that relies on dealer quotes to provide immediacy? What degree of transparency of quotes and trades should be provided? Will traders be able to remain anonymous? How automated should the market be? What should be the minimum tick size at which quotes are made and trades take place? What kinds of orders beyond the standard market and limit orders should be possible?

The answer to these market design questions ultimately depends on how the sources of trading friction are affected and how well the trading needs of investors are met. Will order-processing costs be reduced? Will risk bearing by dealers and/or limit orders be enhanced? Will the problem of free trading options become greater or less? Will the problem of adverse information become greater or less? Will investors be able to

trade quickly? The successful market is one that allows investors to trade when they want to trade, that minimizes real costs of processing orders and of bearing risk, and that deals effectively with the problem of wealth redistribution from informed and speedy traders to uninformed and slow traders.

In this section we first discuss the call auction process. While most markets offer continuous trading, many open with a call auction process. Next the issue of dealer versus auction markets is examined, with particular emphasis on the developments in Nasdaq. Finally a number of other issues in market design are considered.

VI.A. Call auction markets

Call auction process

Most markets began as call auction markets simply because there was not enough activity to warrant continuous trading. Today most markets are continuous. However, the call auction mechanism continues to be used to open trading or to restart trading after a halt.[10] In a call auction market, orders are accumulated and executed at a given time and typically at a single price, p^*, at which supply equals demand. Buy orders at p^* or more buy at p^*. Sell orders at p^* or less sell at p^*.[11]

The benefit of a call market is that it aggregates significant trading interest at particular points in time and limits the free trading option. The free trading option is limited for two reasons. First, since all orders will execute at the auction price, aggressive limit orders can be placed without fear of being picked off at those prices. Second, insofar as the auction is transparent and order may be revised, traders can adjust prices as they see other traders place orders and as they see new information. Adverse information effects may also be reduced in a call auction insofar as investors are able to observe order placement prior to the final price determination. For example, observing a large order to sell will cause potential buyers to adjust their buy orders. Despite the advantages of a call market, most markets are continuous. Investors appear to prefer a continuous market in which they can trade at any time.

---

[10] Markets such as the NYSE, the Tokyo Exchange, and the Deutsche Boerse open with an auction procedure. However, Nasdaq allows each trader to start trading at his quotes.

[11] Because of discreteness in order flow, buy volume need not exactly equal sell volume at a given price. Exchanges establish rules on how such volume is allocated.

It is widely accepted that the most critical and most volatile time in a market's operation is the opening, which typically begins with a call auction. At the opening, information disseminated overnight must be incorporated in securities prices, and orders accumulated overnight must be traded. The final outcome of the opening depends on the net demand of investors and the response of liquidity suppliers.

The working of a call auction market also depends on the rules of the auction. Important issues are the following:

- What degree of transparency exists? Can investors see all orders and the likely opening price? If they can, better inferences can be made about the presence of informed traders.
- Can orders be canceled and revised on the basis of trial opening prices or is this a one shot auction?[12] Disclosure of trial opening prices conveys information and will cause order cancellations and new orders. The ability to cancel orders may also encourage manipulation. One solution is to impose fees for canceling orders and to provide incentives to place orders in a timely fashion.
- Can dealers participate in the auction? On the NYSE, the specialist, and only the specialist, observes the orders and may participate in the auction. This creates a conflict of interest that would not exist if orders were public.

Call auction price determination in the presence of a single monopolistic informed trader is modeled by Kyle (1985) in one of the most cited papers in the field of microstructure. In the Kyle model, the price is determined in a one-shot auction where uninformed investors and the single informed investor place their orders. Trading by the uninformed investors is exogenous and normally distributed with mean zero and variance s_u^2. The informed investor knows the distribution of the uninformed order flow (but not its actual value) and takes account of the impact of his order flow on the market clearing price. The auctioneer determines the auction price to reflect the information contained in the aggregate order flow. Let the asset price before the auction be p_0 and let the variance be s_p^2. Kyle shows that the market clearing price will be

$$\tilde{p} = p_0 + l(\tilde{x} + \tilde{u}), \tag{25}$$

[12] An excellent analysis of a one shot auction is in Ho, Schwartz, Whitcomb (1985).

where $\tilde{x},\tilde{u}$ are the order flow of the informed and the uninformed respectively, and where $l = 2\left[s_p^2 / s_u^2\right]^{\frac{1}{2}}$. The price impact coefficient, λ, is larger the smaller the variance of the uninformed order flow (because it is more difficult for the informed investor to "hide").

Evidence on openings.

Amihud and Mendelson (1987) implement an interesting approach to assessing the volatility around the opening while holding constant the amount of public information released. They calculate daily returns from opening prices, r_0, and from closing prices, r_c. Both returns span a 24 hour period and thus contain the same amount of public information and the same variability due to public information. Stoll and Whaley (1990) apply the Amihud and Mendelson procedure and analyze the sources of volatility around the opening. Based on a sample of 1374 stocks over a five-year period, 1982 – 1986, the average variance ratio is $avg(s_0^2 / s_c^2) = 1.13$. The positive variance ratio implies that opening prices tend to overshoot and reverse after the opening. The reversal of opening prices is reflected in negative serial correlation of open-to-open returns.

Overshooting cannot be ascribed to public or private information arrival because the amount of public and private information is the same for both returns. A possible explanation for overshooting at the open is that trading pressures from liquidity shocks are not completely dampened by liquidity suppliers. Specialists, who are allowed to trade for their own account, may permit prices to deviate from equilibrium in order to earn profits. A second explanation is that the opening is a period of intense price discovery, which requires overnight information to be incorporated in price. The price of a stock is affected not only by the information in the stock but also by information in other stocks. Since all stocks do not open at the same time, some prices must be set in the absence of reliable information as to the value of related stocks. Consequently, some stocks open to high and others open too low. Prices reverse during the trading day as opening pricing errors are discovered. Whatever the exact source of opening volatility, it is an expensive time to trade. Stoll and Whaley compute the Roll implied spread from the serial covariance of open-to-open returns as 0.898% compared with an implied spread of

0.097% from the serial covariance of close-to-close returns. On a 40 dollar stock these implied spreads amount to 36 cents and 3.9 cents, respectively.

In a recent paper, Madhavan and Panchapagesan (2000) compare opening prices in the NYSE opening auction to the opening if the specialist had not intervened. They conclude that specialist intervention is beneficial in bringing the opening price closer to the stock's equilibrium price. In contrast to the NYSE, Nasdaq simply starts trading at posted dealer quotes, which become firm at 9:30 am, the formal start of trading. Cao, Ghysels, Hatheway (2000) analyze the Nasdaq procedure and argue that it works fairly well.

Related to the issue of opening trading is the issue of when to halt trading in a stock or in all stocks. Markets halt trading in individual stocks if news is about to be disseminated or if order imbalances are large. The purpose of such halts is to give investors time to digest the news and determine a new price at which demand and supply are equal. Halts also provide an opportunity for resting limit orders to reset limit prices. In other words, trading is halted in those occasions when re-opening according to a call auction appears desirable. Lee, Ready, and Sequin (1994) analyze trading halts in individual stocks. They conclude that halts have certain benefits, but that volume and volatility increase after a halt. After the crash of October 19, 1987, regulatory circuit breakers were adopted that would shut down trading in all stocks. Currently those circuit breakers are set at 10%, 20% and 30% drops in the Dow Jones Index. A 10% drop shuts the market down for one hour; a 20% drop, for two hours; a 30% drop, for the rest of the day.

VI.B. Dealer versus auction markets: The Nasdaq controversy

In a continuous dealer market, investors buy at a dealer's ask and sell at a dealer's bid. Most bond and currency markets are dealer markets. In a continuous auction market, investors buy at the ask price established by a previously placed sell limit order of another investor and sell at the bid price established by a previously placed buy limit order. Among stock markets, the NYSE is a continuous auction market and Nasdaq is a continuous dealer market, although each has important features of the other. In recent years the Nasdaq Stock Market has come under intense scrutiny and has been required to

undergo major changes. While dealer and auction markets have been the subject of theoretical inquiry,[13] little empirical evidence directly contrasting auction and dealer markets existed prior to the now famous study by Christie and Schultz (1994). Christie and Schultz showed that Nasdaq stocks had a tendency to be quoted in even eighths, necessarily bounding the spread from below at $0.25.

Before presenting some of the evidence on the quality of the Nasdaq and NYSE markets, it would perhaps be useful to contrast the major structural features of these markets:

- The NYSE is centralized exchange where trading takes place on a physical floor (although most orders now arrive electronically), whereas Nasdaq is a physically disperse grouping of dealers each of whom posts quotes on the Nasdaq quotation system.
- The NYSE has 1366 members who must buy seats ($2,000,000 in December, 2000) for the right to trade on the exchange. Seat holders are specialists and floor brokers. Nasdaq has over 5,000 members of whom about 500 are market makers.
- On the NYSE, each stock is assigned to a specialist who makes markets and oversees the book of limit orders. All limit orders on the NYSE are centralized in the book. The best bid and offer, whether for the book or the specialist, are displayed along with the depth at the quote. On Nasdaq, each stock has at least two market makers quoting markets in the stock. The average number of dealers per stock in March 2001 was 11.8, with the top stocks having more than 40 market makers. Each market maker may hold limit orders sent to him and is obligated to display the best bid and offer, whether from a limit order or his own quote, and the associated depth. Prior to the Order Handling Rules implemented by the SEC in 1997, market makers on Nasdaq were not required to display customer limit orders.
- On the NYSE, there has always been a mandated minimum tick size. Until 1997 the minimum increment for quotes and trades was $0.125. On Nasdaq, no increment was mandated, but convention frequently led to trades at increments of even eighths as

[13] See Garbade and Silber (1979), Cohen, Maier, Schwartz and Whitcomb (1981), Ho and Stoll (1983), Madhavan (1992), Pagano and Roell (1992), Biais (1993) and Laux (1995).

found by Christie and Schultz (1994). Under SEC urging, the minimum tick size on the NYSE and Nasdaq was reduced to one cent in 2000.

- On the NYSE, orders may be routed electronically over the DOT system directly to the specialist. Execution is not automatic, but occurs only when the specialist accepts the trade. On Nasdaq, orders may be routed to a market maker electronically over SelectNet, which like the DOT system, requires the market maker to accept the trade. Orders may be automatically executed over Nasdaq's SOES (small order execution system) up to the market maker's posted depth.

Christie and Schultz (1994) investigated the spreads of 100 Nasdaq stocks in 1991 in comparison to 100 NYSE stocks. They find a nearly total avoidance of odd eighths quotes for 70 of the 100 Nasdaq stocks and a resulting higher spread on Nasdaq than on the NYSE. They conclude that Nasdaq market makers are implicitly colluding to keep spreads high. Huang and Stoll (1996) compare execution costs for 175 Nasdaq stocks to execution cost for a matched sample of NYSE stocks in 1991. They find that execution costs as measured by the quoted spread, the effective spread (which accounts for trades inside the quotes), the realized spread (which measures revenues of suppliers of immediacy), or the Roll (1984) implied spread, are twice as large for a sample of NASDAQ stocks as they are for a matched sample of NYSE stocks. The results are in Table 3.

Huang and Stoll (1996) conclude that the higher trading costs in Nasdaq are not due to asymmetric information because the asymmetric information component of the spread, measured as the difference between the effective and realized spreads, is the same in the two markets. Partial explanations are provided by differences in the treatment of limit orders and commissions in the two markets. In Nasdaq, limit orders were not displayed (as are limit orders on the NYSE) and consequently, limit orders could not narrow the spread. In Nasdaq institutional investors pay no commissions, although individual investors do. Thus in the case of institutions some of the difference in spreads in the two markets reflects the fact that NYSE spreads can be lower by the amount recovered in commissions. Huang and Stoll also conclude that spread differences are not related to differences in market depth or in the frequency of even eighth quotes, once stock characteristics are held constant.

Two features of Nasdaq contributed to a lack of competition. First, a common feature of multiple dealer markets is that each dealer seeks to capture a certain fraction of the order flow by internalizing trades from a parent broker or by arranging for trades to be preferenced to it. Internalization occurs when a retail broker sends its order flow to its affiliated dealer. Preferencing occurs when a retail broker arranges to send its order flow to chosen dealers, often in return for a payment. The dealer receiving internalized or preferenced order flow promises to trade at the best quote even if he is not currently posting the best quote. When a large fraction of order flow is preferenced or internalized, little incentive exists for any dealer to compete by narrowing the spread because a large fraction of the order flow is already allocated to other dealers. Indeed, narrowing the spread reduces the revenues of all dealers (because they promise to match the best price) and generates considerable pressure from all dealers not to narrow the spread. A second market structure feature that inhibited competition in Nasdaq was the availability of alternative electronic markets where a dealer could offer better prices to even out inventory without making those prices generally available.[14] Dealers could use Instinet, a proprietary trading system or SelectNet, a Nasdaq system, to trade with other dealers at favorable prices without offering those prices to their retail order flow.

After extended investigations by the Securities and Exchange Commission (SEC) and the Department of Justice, the SEC in 1997 put into effect order handling rules that required limit orders to be displayed and to be given price priority. Strict time priority across dealers and markets is not required. The effect of this rule was to allow limit orders more effectively to compete with dealer quotes. Second, the order handling rules prohibited a dealer from quoting in Nasdaq at a price inferior to the dealer's quote in an electronic communications system (ECN). If the ECN displayed its best quotes in Nasdaq, the dealer obligation to quote the best price in Nasdaq was satisfied. This ECN rule made available to the public the same quotes previously available only on the interdealer market.

[14] Preferencing and the use of inter-dealer trading systems are also common on the London Stock Exchange. Papers by Hansch, Naik, and Viswanathan (1998) and Reiss and Werner (1998) analyze this market and find that there is some price competition and some response of order flow to prices. Wahal (1997) finds that dealer entry is related to spreads, but entry may simply divide the profits among more players without reducing overall profits.

The order handling rules had a dramatic effect on quoted spreads, which fell by 30%, as chronicled in Barclay et al (1999). Effective spreads also fell but not as much. Recent evidence [U.S. SEC (2001] suggests that effective spreads on Nasdaq continue to exceed those for comparable NYSE stocks.

The benefit of a dealer market arises from the flexible response of dealers to liquidity needs. Dealers are able to respond quickly to changing market conditions. Yet evidence indicates that dealers, left to themselves, raise spreads above those observed when limit orders are also displayed. The benefit of an auction market is that limit orders from the trading public provide liquidity. Fischer Black (1971) predicted that an automated market (much like the new ECNs) would be able to operate without dealers, and that dealers would be driven out of business. It does not appear, however that a pure limit order market is able to provide sufficient liquidity, particularly in less active stocks. Dealer intervention is often needed to bridge gaps in the arrival of limit orders. On the NYSE, for example, the specialist participated on the buy or sell side in 27.5% of the share volume in 2000.[15] The implication is that a mixed dealer/auction market is optimal.

VI.C. Other issues in market design

Market centers face a number of other design issues, including the degree of transparency, whether traders remain anonymous, whether trading is fully automated, what minimum price increment should be established, and the kinds of orders that are allowed.

Transparency

Transparency refers to the disclosure of quotes (at which trades can take place) and of transaction prices (at which trades did take place). The NYSE displays only the top of the book, that is the best bid and ask, but not the other orders on the book. The ECNs display the entire book. The benefits of transparency are three-fold. First, transparency speeds price discovery and enhances market efficiency, for with transparent markets all investors see the current quotes and the transaction prices, and no investor trades at the wrong price. Second, transparency helps customers monitor brokers. The public dissemination of quotes and transactions allows a customer to determine that his

[15] NYSE, *Fact Book, 2000 Data*, p.18.

transaction is in line with others at the same time. Third, transparency enhances competition, for it allows competing dealers to guarantee the best price anywhere, but do it at a lower commission or lower spread.[16] The costs of transparency arise from adverse incentive effects. First, traders may be reluctant to place limit orders, particularly if they are large, because the display may convey information that will make the price move against the limit order. Second, display of limit orders may make it easier for traders to exercise the free trading option and thus reduce the incentive to place limit order. If no one knows whether a limit order exists, it is more difficult to pick it off, but if the limit order is displayed, it can be more readily picked off.

Anonymity

Closely related to the issue of transparency is the issue of anonymity. Should the identity of traders be known? Some traders, such as dealers want to be identified because the want to build reputations. Other traders, such as institutions who are likely to be informed, want to be anonymous because disclosure of their identity may cause prices to move against them. If they cannot capitalize on their special information, their incentive to do research is reduced, and information production could be harmed. Admati and Pfleiderer (1991) analyze the idea of sunshine trading by which an uninformed investor creditably reveals himself and thus prevents an adverse price reaction. Several papers, including Benveniste, Marcus and Wilhelm (1992) and Forster and George (1992), analyze the effect of anonymity.

Automation

Automation is an issue because it affects the value of the free trading option and who has it. When execution is automatic at a dealer's quote, the dealer grants the option. Furthermore, if the dealer is slow to update quotes, several trades might take place before the quote can be changed. The SOES (Small Order Execution System) system of Nasdaq worked in this manner. Upstairs traders sitting at terminals often placed orders more quickly than the reaction time of the dealer. In an order routing system like the NYSE DOT system or the Nasdaq SelectNet system, orders are delivered to the dealer, but the dealer must accept the order within a specified period of time. This gives the dealer some time to react and perhaps change the quote. In effect the dealer now has the option.

[16] Madhavan (1995) analyses the effect of transparency on fragmentation and competition.

Before automated routing and execution systems, orders were hand-carried to the floor and some negotiation took place. A completely automated system does not permit negotiation. Hence, a completely automated system is more successful for orders that do not require negotiation, such as most small orders. Large orders, where negotiation is common, are not automated (except in so far as a computer system mimics a negotiation). In an interesting theoretical paper Glosten (1994) shows that an open electronic limit order book would be most efficient and would dominate other exchanges.

Tick size

The tick size is the minimum allowable price variation in a security, usually determined by the exchange on which the security trades. On the NYSE, the tick size before June 1997 was $1/8^{th}$ dollar by rule. In futures markets, each futures contract has a specified tick size that depends on the value of the futures contract and its variability. For example, the tick size for the S&P 500 futures contract is 0.10 index points or $25 per contract. Mandated tick sizes are not common in dealer markets. For example, Nasdaq has not had a market-wide mandated tick size, although convention led to a minimum tick size of $1/8^{th}$ with a number of stocks trading at wider increments as discussed above. Under SEC pressure, the tick size in U.S. equities markets was reduced to $1/16^{th}$ in the 1997 and to one penny in 2001.

The tick size has several effects. First, the tick size affects incentives to place limit orders, as Harris (1991) first noted, since it represents the cost to getting inside someone else's quote. If the tick size is 12.5 cents, and the standing bid is $20, one must bid at least 20.125 to move ahead of the standing bid. If the tick size is one cent, one must bid only 20.01 to move ahead of the standing bid. Since it is easy to move ahead of a limit order when the tick size is small, fewer limit orders will be placed when the tick size is small, which can have adverse effects on liquidity. A second effect is that a mandated tick size can cause spreads to be artificially large, at least for some trades. [17] When the tick size is 12.5 cents, the minimum spread is 12.5 cents. A 12.5 cent spread may exceed the equilibrium spread for 100 share orders, causing such orders to pay too much.

[17] See Harris (1994). Hausman, Lo and MacKinlay (1992) and Ball and Chordia (2001) provide approaches to analyzing true price behavior and true spreads in the presence of artificial price increments imposed by the minimum tick size.

Currently with a tick size of one cent, many stocks trade at a spread of 5 cents or less, but the depth is less than it would be at a 12.5 cent spread. When the tick size is small and depth at the inside quote is small, it is important that markets display information on the available liquidity at prices away from the inside quote in order to give investors information as to the likely price at which they can trade their orders.

Order types

Another issue in market design is the types of orders that will be allowed. On the one hand a market may wish to restrict certain common order types. For example, electronic markets often forbid straight market orders, requiring instead the use of marketable limit orders. A market order would execute at any price. If the book is thin and another order takes the quantity displayed at the inside, an unsuspecting market order might trade at prices far removed from the equilibrium price. A marketable limit order is an order at the current market price that pays no more (or receives no less) than the current price. On the other hand, automated exchanges offer the possibility of much more complex order types. For example, contingent orders could easily be monitored in a computer. A contingent limit order that adjusts the limit price based on the price of the stock or an index can reduce the free trading option and can alleviate the chance that a limit order is picked off. Opponents of automatic quote updating fear that markets will become computer dueling grounds in which traders program their order submission strategy, turn on the computer, and go back to bed. Nasdaq, to limit pure computer trading, has limited the ability of dealers automatically to update quotes.

VII. The market for markets: centralization versus fragmentation of trading

Trading of stocks and related instruments takes place in a variety of different markets. Stocks listed on the NYSE trade there, but also trade on regional exchanges, in the third market, and on some other proprietary systems. Trading of stocks listed on Nasdaq trade there, but also trade on ECN's and on other proprietary systems. Many U.S. stocks trade in foreign markets. Options on stocks trade in five option markets. Futures markets trade stock indexes, and have recently received regulatory approval to trade futures on individual stocks. While the number of markets existing today is greater than ever in the past, many observers argue that markets will merge and consolidate, while

others predict increased fragmentation of markets. In this section the evolution of U.S. equities markets and of global equities markets in the last 30 years is reviewed and the forces of centralization and of fragmentation are discussed.

VII.A Evolution of U.S. equities markets.

In 1970, the New York Stock Exchange (NYSE) accounted for the overwhelming bulk of trading in stocks, and it faced little or no competition. The American Stock Exchange (AMSE) did not compete because, by agreement, it listed and traded only stocks not listed on the NYSE (accounting for 11% of dollar volume of all listed stocks). The Nasdaq Stock Market did not yet exist, although stocks that were not yet eligible for listing were traded on the OTC market. Stocks listed on the NYSE could be traded on regional stock exchanges under an SEC rule that granted them unlisted trading privileges (UTP). The regional exchanges (Midwest, Pacific, Philadelphia, Boston, Cincinnati) accounted for only 12% of dollar volume of stocks listed on the NYSE. The organization of the NYSE met the classic definition of a cartel:

- limited membership --one must own one of 1366 seats in order to trade on the NYSE,
- fixed prices -- commission rates were fixed,
- rules and regulations limiting non-price competition among cartel members –price discounts were prohibited, and Rule 394 prohibited members from trading off the NYSE where they could charge lower commissions.

By 2000, the organization of trading markets had changed in response to technology and regulation. Fixed commissions were abolished in May, 1975. The Nasdaq Stock Market, founded in 1971, now rivals the NYSE with dollar volume exceeding that on the NYSE.[18] With the growth in Nasdaq, the AMSE lost its second place position as a stock market. Instead it has become an index and options market. The regional exchanges (Boston, Cincinnati, Midwest, Pacific, Philadelphia), despite predictions of their imminent demise, have maintained their overall share of NYSE dollar volume, but they continue to be under pressure. A host of new proprietary trading systems that include Instinet, a system aimed at institutional traders, and other electronic communications

[18] Part of this reflects the trading system of Nasdaq where a dealer tends to be involved as both a buyer and seller, whereas on the NYSE customer to customer trades are more likely.

systems (ECNs) that totally automate trading, now compete for order flow. Some of the major features of changing market organization are outlined here.

Competitive commissions

In 1970, commissions on a 500 share trade of a 40 dollar stock were $270 dollars.[19] While institutional investors received a quantity discount, they still paid substantial amounts (for example, 26.2 cents per share on a 5000 share trade). Economic pressures on commissions took two forms. First dealers outside the NYSE offered to trade shares at discounted commissions. The **third market** is the market in NYSE stocks made by brokers and dealers who are not members of the NYSE (and thus exempt from rule 394). In the 1960s and 1970s, institutional investors used the third market to reduce commissions. Second, while NYSE rules limited rebates of commissions, they did not limit service competition. Consequently brokers rushed to provide services and products in return for lucrative commission business. **Soft dollars** are that portion of the commission over and above the cost of doing the trade. Institutions paid soft dollars for research services, mutual fund sales, phone lines, and a variety of other services. Soft dollars still exist today, but they are limited by regulation to research services, and the amounts are smaller.

In addition to the economic pressures on commission, the Department of Justice and the SEC also attacked fixed commissions. Finally, Congress abolished fixed commissions as part of the Securities Acts Amendments passed on May 1, 1975. Dire consequences were predicted, but the securities industry easily survived the change, as reductions in commissions were more than offset by increased trading volume and more efficient trading procedures.[20] Today the cost of a 500 share trade, handled electronically, is typically less than $25 (despite the inflation since 1970), and institutions typically trade for 5 cents per share.

Rule 394 and the third market

[19] See Stoll (1979) p. 13.
[20] The effects of the May Day 1975 changes are analyzed in Stoll (1979).

In 1970, NYSE Rule 394 prohibited member firms of the NYSE from trading outside the NYSE either as agent or as principal. Member firms, acting as agents, could not send customer orders to other markets (other than regional exchanges), nor could they trade with customers as principals outside the NYSE. This rule had the beneficial effect of forcing all orders to interact in one market – the NYSE, but it had the harmful effect of limiting competition from new markets. Over time, regulatory pressure weakened Rule 394 and caused it to be abolished in 2001. First, in 1976, the rule was changed to Rule 390, which permitted trades, where the NYSE member acted as agent, to be executed off the NYSE. This modification gave rise to a new third market as member firms sent customer orders to third market makers (such as Madoff and Co.) that promised to match NYSE prices. In addition the third-market-maker paid the broker for the order flow. The new third market specialized in the order flow of small, uninformed, customers in contrast to the third market of the 1970s, which was an institutional market to avoid high commissions.

Second, Rule 390 was weakened by SEC Rule 19c-3 that exempted any stocks listed after April 1979 from application of the rule. Under 19c-3, a NYSE member could trade with customers as a principal and could therefore make in-house market in eligible stocks, but, surprisingly, few members set up in-house markets in listed stocks. Finally, Rule 390 was abolished by the NYSE in 2001 because of SEC pressure and because the rule had become ineffective.

Thus by the year 2001, two of the key features of the NYSE cartel – fixed commissions and the restrictive Rule 394 – had been abolished. The one remaining feature of the cartel – limited direct access for the 1366 members – remains. The privilege of membership continues to have substantial value as NYSE seat prices in 2000 exceeded $2 million. Members are of three types:[21] specialists (about 450), independent floor brokers (about 525) and floor brokers for retail firms (about 330). Specialists trade for their own accounts as market makers and keep the book of limit orders. Independent floor brokers receive commissions for executing customer orders. Floor brokers that work for retail firms execute the portion of the firms' order flow that is not routed through the electronic DOT system.

[21] See Sofianos and Werner (2000) for a description of the membership.

National Market System

The Securities Acts Amendments of 1975, in addition to abolishing fixed commissions, directs the SEC to facilitate the establishment of a "national market system" that is characterized by the absence of unnecessary regulatory restrictions, fair competition among brokers, dealers and markets, the availability to all of information on transaction prices and dealer price quotations, the linking of markets and the ability to execute orders in the best market. The SEC envisaged a single national market in which orders would be routed to the best market and in which a single CLOB (consolidated limit order book) would contain limit orders and dealer quotes in each stock. A single CLOB has not been implemented, as it would require substantial integration of different markets and would limit competition.

Certain elements of an NMS have been introduced. These include the consolidated trade system (CTS), the consolidated quote system (CQS), and the intermarket trading system (ITS). CTS and CQS enhance market transparency as they require all exchanges to report centrally their transactions (price and quantity) and quotes, and thereby enable traders in any market to determine if they are trading at the best prices. The CQS and CTS do not provide access for brokers and dealers on one floor to better quotes on another floor. Access is provided through ITS, which links exchange floors and permits traders on one floor to send a "commitment to trade" to another floor. The other floor has a limited time to accept or reject this commitment.

The future of the national market system is cloudy. On the one hand some observers argue that the SEC should impose tighter links among markets and improve ITS. On the other hand, some would let the nature and extent of links be decided by markets and by investors on the basis of available technology. In fact, computer routing systems can quickly send an order to the best market, without the need for a government sponsored CLOB or ITS.

VII.B. Global markets

Equities markets in other parts of the world have changed as much and as rapidly as U.S. markets. In October 1986, the London Stock Exchange (LSE) underwent the "big

bang" by which fixed commissions and a restrictive jobber system were eliminated and a dealer trading system similar to Nasdaq was adopted. In the late 1980s, Paris replaced its floor trading system with a computerized limit order book, which is analyzed in Biais, Hillion and Spatt (1995). Toronto was an early adopter of a computerized trading system in 1977. The German markets were late to change but have done so with a vengeance. The Deutsche Boerse is a for-profit business overseeing the automated stock trading platform, Xetra, and several other markets, including the electronic futures market, Eurex. A merger between the LSE and the Deutsche Boerse was attempted but failed. The Paris Bourse has successfully consolidated with Amsterdam and Brussels to form Euronext. As in the U.S., private electronic trading systems are also making inroads in Europe. Domowitz (1993) provides a comparison of automated trading systems around the world.

As markets around the world develop, they are in a position to trade securities from any other part of the world. As a matter of technology, the stock of an American company can be traded as easily on the LSE as on the NYSE. However, globalization of markets has not proceeded as rapidly as technology allows. Stocks domiciled in the U.S tend to trade primarily when U.S. markets are open and stocks domiciled in Europe tend to trade primarily when European markets are open. There is evidence of some migration of trading from one country to another in the same time zone [Domowitz, Glen and Madhavan (1998).

Cross-listing of stocks from one country on the exchange of another country is often done in the form of depositary receipts. In the U.S., American Depositary Receipts (ADRs) are dollar denominated claims issued by a bank on the underlying shares held by a bank. For example, British Telecom ADR traded on the NYSE is a claim on 10 shares of British Telecom traded in London. Arbitrageurs keep prices of the ADR and UK shares in equilibrium. Nothing in principle prevents stocks from being listed in the U.S. in terms of their home currency. Traders of such shares in the U.S. must have the ability to pay or receive a foreign currency. Alternatively, nothing prevents a company from listing its shares in a variety of countries in terms of each local currency.

One of the puzzles in international finance is the slowness with which international diversification has taken place. Investors are said to have a home bias.[22] This phenomenon is reflected in the slowness with which stocks are traded internationally. Stock trading for most companies is concentrated in the company's home country by those investors domiciled in that country.

VII.C. Economic forces of centralization and fragmentation

In spite of the weakening of the cartel rules of the NYSE, the NYSE continues to attract most of the order flow in the stocks it lists. At the same time new markets are being founded almost daily both in the U.S. and abroad. Consequently there is a tension between centralization of trading in a single market and the initiation of new markets that fragment trading. Fragmentation of trading can be said to arise when an order in one market is unable to interact with an order in another market.

The forces of **centralization** are two-fold – one on the supply side and one on the demand side. First, on the supply side, a market reaps economies of scale in processing transactions. The average cost of trading a share of stock declines with the number of shares traded. As a result, the first mover into the trading business has a great advantage because it can process trades at lower cost than a competitor using the same technology. Second, on the demand side, a market generates network externalities. A market is a communications network, and like other networks, its attractiveness depends on the number of others on the network. Traders want to trade where other traders are already trading because the probability of a successful trade is a function of the number of other traders using the market. Consequently, network externalities, like economies of scale, lead to a first mover advantage.

Several factors have made **competition** from satellite markets more effective in recent years and have weakened the centralizing forces of economies of scale and network externalities. First, the transparency of quotes and transaction prices makes it possible for a satellite market to credibly guarantee that the price in the primary market is being matched. For many years, the NYSE jealously guarded its price information and

[22] For example see Cooper and Kaplanis (1994), Telsar and Werner (1995), Kang and Stulz (1997) and the chapter in this volume by Karolyi and Stulz..

limited the dissemination of its quotes and transaction prices. Without knowledge of where the price is, investors prefer the primary market where price discovery takes place. With transparency, a trader can be assured the price in a satellite market at least matches the price in the primary market.

Second, satellite markets not only match prices, but they also pay for order flow from brokers. A typical payment might be one or two cents per share for market orders from retail investors that are judged to be uninformed. Payment is not made for limit orders or for order flow judged to be informed. Payment for order flow has been criticized because the payment goes to the broker, not to the customer whose order is being routed to the satellite market. While payment for order flow is quite common among satellite exchange, it is not necessarily sufficient to overcome the natural centralizing forces. If the primary market is the low cost producer of transaction services, it can make the same payment.[23]

Third, technological change has made competition more effective. Nimble new exchanges may be able to implement new, low cost, electronic trading systems more quickly than existing markets and thereby attract order flow away from established markets. Communications technology also reduces the switching costs of moving trades from one market center to another. The ease with which orders can be routed to a satellite market has improved.

Fourth, regulatory policy in the U.S. has fostered competition and fragmentation. The SEC has required greater transparency, which enhances competition from new markets. Second the SEC has required markets to link, which has given satellite markets access to the primary market. Such links enable dealers in the satellite market to lay off inventory in the primary market and provide an opportunity for brokers to route orders to the satellite market.

It is not evident how the conflict between centralization and fragmentation will be resolved in the future. The forces of centralization – economies of scale and network externalities – are strong. While they have been weakened by technology and regulation, they have not been weakened to the extent that markets will necessarily fragment into

[23] Battalio, Greene and Jennings (1997) conclude that preferencing arrangements on the Boston and Cincinnati stock exchanges attracted order flow to those exchanges without adversely affecting the quality of markets.

many separate unconnected market centers. If markets do fragment, the adverse consequences are small because markets are linked by high speed communications systems. The term "fragmentation" has a harmful connotation, but, in fact, fragmentation is just another word for competition. Competition among markets is a good thing because it fosters innovation and efficiency. Separate markets may exist, but when linked by high-speed communications systems they act almost as one.

The cost of fragmentation is that priority rules are difficult to maintain across markets. Price priority can usually be maintained because, with transparency, the investor can send his order to the market with the best price. But even price priority can sometimes be violated, for example, when large orders in one market trade through prices in another market. Time priority is likely to be violated as traders prefer to trade in one market over another that may offer the same price.

VII.D. The future structure of markets[24]

The evolution of the securities industry will be shaped by technology and by regulation. Technology widens the extent of the market beyond a particular region or a particular country. Communications technology links investors to all markets and hence intensifies competition among existing market centers. Foreign markets can easily trade U.S. stocks, and U.S. markets can easily trade foreign stocks.

Technology changes the nature of exchanges. In the past securities were traded on membership exchanges – mutual organizations organized more like clubs than like businesses. However, the task of trading securities has become a business with private firms taking a larger role. As a consequence some exchanges have de-mutualized in an attempt to organize themselves more effectively and with an eye to raising capital by stock sales.

Technology changes the relative position of customers, retail brokers, exchanges, and market making firms. Retail firms and customers have the ability to create their own markets and put pressure on exchanges to respond to their interests. Large national market making firms are able to trade their order flow on any of a number of markets,

[24] For some thoughtful predictions, see Lee (2002)

thereby put competitive pressure on exchanges. New electronic markets provide low cost trading and put pressure on existing exchanges.

Regulation sets the rules for competition among market centers. The SEC has pushed for links among markets and transparency of prices and quotes. By and large this policy has enhanced competition, but it has limited the flexibility and speed with which markets could act. SEC rules recognize that all market centers are not equal. The SEC rule on alternative trading systems (ATS) distinguishes exchanges and ATS. ATS are electronic trading systems that do not carry out all the functions of an exchange. ATS are regulated as broker dealers with additional requirements depending on their size. An exchange has self-regulatory obligations, has requirements as to governance and board structure, and must participate in market linkages. While exchanges sometimes criticize the SEC for imposing on them the costs of regulating their markets, SRO responsibilities often become a competitive advantage vis a vis non-exchange market centers. In addition, exchanges reap substantial revenues from the sale of quote and price information. As a consequence, several ECNs have applied for exchange status.

VIII. Other Markets

Market microstructure research has focused on equities markets, but other market are clearly important, albeit, less studied.

Bond market

The bond market is a dealer market. Dealers display indicative quotes and provide firm quotes in response to customer inquiries. Customers trade directly with dealers, at dealer prices. Dealers can trade anonymously with other dealers through inter-dealer brokers. Inter-dealer brokers display anonymous dealer quotes, usually only to other dealers, and execute inter-dealer transactions.[25] Participants in the bond market are institutional investors – insurance companies, investment companies, banks, etc. – who trade in relatively large amounts. Individual investors are not a major element in bond trading. Secondary market trading of bonds is relatively infrequent as the bonds are often held to maturity.

[25] Exclusive inter-dealer trading also existed in the Nasdaq Stock Market, but was eliminated by the SEC on the grounds that this was a mechanism that contributed to high bid-ask spreads for the public.

The microstructure of bond markets has not been studied to the same extent as the microstructure of equities markets, partly because data are not readily available. An early study by Fisher (1959) showed that corporate bond yields varied by marketability. More marketable bonds (measured by number of bonds outstanding) are priced at lower yields to maturity. Grant and Whaley (1978) show that bond spreads depend on risk as measured by duration as well as on quantity outstanding. Hong and Warga (2000) compare transactions data from the NYSE Automated Bond System, where transactions are of small size, and from insurance companies, which trade in large size, and conclude that effective spreads from these two sources are quite similar. Schultz (2001) examines the quoted bid-ask spread of corporate bonds as a function of bond characteristics. He concludes that trading costs are lower for larger trades, which reflects the fact that the bond market is largely an institutional market.

The most active bond market is that for U.S. treasuries, and the most active time is at the initial offering of bonds. Unlike stock issues that occur at a given offering price, bonds have been issued in a sealed bid price discriminatory auction. Jegadeesh (1993) studies Treasury auctions in the period 1986 – 1991. He finds that the "on the run" bond is typically priced above comparable bonds in the secondary market and that the bid-ask spread is below that in the secondary market. Secondary market trading of government bonds is studied by Elton and Green (1998). They find that most of the cross sectional variation in bid-ask spreads can be explained by factors such as volume and maturity. However they conclude the effect of liquidity on bond prices is small.

Currency market

The currency market is a dealer market made largely by the same dealers active in the bond market. Currency dealers display indicative quotes, but quotes at which trades may occur are usually made bilaterally. Like the bond market, the currency market has an interdealer market in which dealers can trade anonymously with each other. Lyons (1995) analyses the behavior of a major currency dealer and concludes that inventory considerations are important determinants of dealer behavior in two senses. First, there is a direct effect from the dealer's desire to have a zero position at day-end. Second, there is an indirect effect from information about other dealers' inventories that influences the dealer's behavior. Other articles examining the microstructure of currency markets

include Bessembinder (1994), Bollerslev and Melvin (1994) and Huang and Masulis (1999).

Futures markets

Organized futures markets are open outcry auction markets. Trading takes place in a pit where traders, representing themselves or customers, signal their desire to trade. In major contracts such as index futures or T-bond futures, hundreds of traders are present, and trading is extremely rapid. Many transactions may occur at nearly the same time. Consequently, unlike the equities market in which all quotes and transaction prices are reported sequentially, in the futures markets not every quote nor every transaction is reported on the ticker tape. Liquidity is provided by scalpers who buy contracts at their bid price and sell contracts at their ask price. Manaster and Mann (1996) analyze floor traders in futures markets. They find evidence of inventory management in that inventories are mean reverting. On the other hand they also find that traders do not pay price concessions in order to manage their inventory.

Options markets

Active secondary markets in options on common stocks date to the founding of the Chicago Board Options Exchange (CBOE) in 1973. Today equity options are traded on the CBOE, on three other traditional exchanges (American, Philadelphia and Pacific), and on a new electronic exchange, the International Securities Exchange (ISE). The American and Philadelphia exchanges employ a single specialist system whereas the CBOE and Pacific Exchange use a competing market maker system. In recent years the CBOE and Pacific have designated primary market makers and given these firms more responsibility for overseeing the markets in their options. As a practical matter, option trading is more complicated than stock trading simply because the large number of different option contracts for any given stock. For example, IBM stock has over 100 puts and 100 calls with different maturities and strike prices. When the stock price changes, all the option prices must be updated quickly.

The microstructure of options have been analyzed from a number of different perspectives. Vijh (1990) examines option spreads and the price impact of large options trades. He concludes that large options trades are absorbed well by the market. Spreads are as large as those in the underlying stock despite the lower price of the option. A

number of papers have investigated whether options prices lead prices of the underlying stocks [Stephan and Whaley (1990), Chan, Chung and Johnson (1993), Easley, O'Hara and Srinivas(1998)]. One would expect such a lead if the informed investors trade in the more leveraged options market rather than in the stock market, but the evidence is mixed.

IX. Asset pricing and market microstructure.

It seems obvious that microstructure factors ought to affect asset prices. Consider for example a firm raising equity capital for the first time. The price investors would pay for the new shares must undoubtedly depend on the ease with which those shares can be sold in the future. If all investors face a cost of selling the shares that is 20% of the price, the value of the shares will certainly be much lower today than if the disposition cost were 2%. The valuation effect of real friction, such as the cost of processing orders or searching for counterparties, is clearly to reduce an asset's value. The valuation effect of informational friction is less clear. Informational friction arises if one investor is better informed than another. The informed investor with good news will bid up asset prices to the disadvantage of the uninformed investor who sells the shares. Similarly, when disposing of shares, the informational investor receives a better price than the uninformed investor. The presence of informed investors disadvantages uninformed investors and redistributes income from the uninformed to the informed. Informational frictions introduce distributional uncertainty, which may make some investors reluctant to buy an asset, thereby lowering its market price.

A number of studies have examined the relation of microstructure and asset pricing. Stoll and Whaley (1983) show that expected returns are related to transaction costs and they argue that the small firm effect can be explained at least in part by the higher transactions costs of small firms. Amihud and Mendelson (1986) develop and test a model of asset pricing with transaction costs. Brennan and Subrahmanyam (1996) show that required returns are related to the Kyle price impact coefficient. Brennan, Chordia and Subrahmanyam (1998) show that expected returns are negatively related to volume after controlling for other factors such as firm size and book to market ratio, a result they attribute to greater liquidity and lesser trading costs of high volume stocks.

X. Conclusions

In the past twenty years, research on the simple question of what happens when financial assets are bought and sold has grown to the extent that it is now a recognized sub-field within finance -- market microstructure. Probably the field has grown so dramatically simply because it is interesting. Microstructure research examines the process of price formation in the presence of risks, costs and asymmetric information, factors that are central to finance. Add to that the availability of large transaction data bases, and one has a recipe for a successful research area.

Microstructure research has also grown because the field deals with important practical issues. Microstructure research influences regulatory policy, such as the regulation of the Nasdaq Stock Market. Microstructure research contributes to institutional trading strategy and the proper measurement and management of trading costs. Microstructure research provides an intellectual framework for designing and operating trading systems.

In this chapter I have tried to convey some of the important institutional features of markets while also presenting the ideas that underpin the scholarly study of market microstructure. Scholarly analysis focuses on the determinants of the bid-ask spread and on the effect of market frictions for short-term behavior of asset prices. If there were no market frictions, bid and ask prices would be equal, and short-term price fluctuations would depend only on information arrival. In fact, market friction, resulting from the costs of processing orders, from inventory risk assumed by suppliers of liquidity, from free options granted by liquidity suppliers, and from asymmetric information, lead to differences in bid and ask prices and to short term price volatility. A desirable market design is one that minimizes the effect of these trading frictions. Evidence suggests that continuous markets are preferred to call markets and that a market that combines features of dealer and auction markets is superior to a pure dealer or auction market.

Markets experience economies of scale and network externalities that could lead to domination by one market, but competition is desirable because it encourages innovation and efficient market design. In recent years, a variety of new markets have challenged established markets with the result that no exchange has achieved a level of dominance that would be implied by economies of scale and network externalities. We

can ascribe the competition among markets to the transparency of market price information that enables satellite markets to match prices in the primary markets, to regulatory action, and to innovations by new markets to provide trading technology or appeal to niches of the market not well served by the primary market.

Microstructure remains a fertile field for additional research. The field has focused on relatively narrow questions with little attention to its implications for broader issues such as asset pricing. How precisely and to what degree do measures of liquidity affect asset pricing? To put it another way, the relation between microstructure of financial market and the macrostructure of financial markets deserves further study.

Within the narrower confines of the microstructure sub-field, a variety of issues remain to be resolved. For example, it is not yet clear which -- asymmetric information, inventory or order processing costs -- are the most important factors in the bid-ask spread. Nor is it clear how these components vary across stocks or how they are affected by regulation, by market design and by stock characteristics. What is the relation between different measures of liquidity? Is the spread of a stock a good predictor of the price impact that might be caused by a trade? These and related questions should keep researchers busy for a while.

References

Admati, Anat R., and Pfleiderer, Paul. 1988. A theory of intraday patterns: volume and price variability, *Review of Financial Studies* 1, 3-40.

Admati, Anat R., and Pfleiderer, Paul. 1991. Sunshine trading and financial market equilibrium. *Review of Financial Studies* 4 (3), 443-481

Amihud, Yakov, and Haim Mendelson, 1980, Dealership market: Market making with inventory, *Journal of Financial Economics* 8, 31-53.

Amihud, Yakov, and Haim Mendelson, 1986, Asset pricing and the bid-ask spread, *Journal of Financial Economics* 17, 223-249.

Amihud, Yakov, and Haim Mendelson, 1987, Trading mechanisms and stock returns: An empirical investigation, *Journal of Finance* 42, 533-553.

Bagehot, Walter (pseudonym for Jack Treynor), 1971, The only game in town, *Financial Analysts Journal* 27, 31-53.

Ball, Clifford, and Tarun Chordia, 2001, True spreads and equilibrium prices, *Journal of Finance* 56: 1801-1835.

Barclay, Michael, William Christie, Jeff Harris, Eugene Kandel, and Paul Schultz, 1999, The effects of market reform on the trading costs and depths of Nasdaq stocks, *Journal of Finance* 54, 1-34.

Battalio, Robert, Jason Greene and Robert Jennings, 1997, Do competing specialists and preferencing dealers affect market quality? *Review of Financial Studies* 10, 969-993.

Benston, George, and Robert Hagerman, 1974, Determinants of bid-asked spreads in the over-the-counter market, *Journal of Financial Economics* 1, 353-364.

Benveniste, L., A. Marcus, and W. Wilhelm, 1992, What's Special about the Specialist?" *Journal of Financial Economics* 32(1) 61-86.

Bessembinder, Hendrik, 1994, Bid-ask spreads in the interbank foreign exchange markets, *Journal of Financial Economics* 35, 317-348.

Biais, Bruno, 1993, Price formation and the equilibrium liquidity in fragmented and centralized markets, *Journal of Finance* 48, 157-186.

Biais, Bruno, Hillion, Pierre and Chester Spatt, 1995, An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse, *Journal of Finance* 50 (December) 1655-1689.

Black, Fischer, 1971, Toward a fully automated exchange, *Financial Analysts Journal, Part I*, (July/August), *Part II*, (November/December).

Bollerslev, T and Michael Melvin, 1994, Bid-ask spreads and volatility in the foreign exchange market: an empirical study, *Journal of International Economics* 36, 355-372.

Branch, Ben, and Walter Freed, 1977, Bid-ask spreads on the AMEX and the Big Board, *Journal of Finance* 32, 159-163.

Brennan, Michael J., and Avanidhar Subrahmanyam, 1996, Market microstructure and asset pricing: On the compensation for illiquidity in stock returns, *Journal of Financial Economics* 41, 441-464.

Brennan, Michael J., Tarun Chordia, and Avanidhar Subrahmanyam, 1998, Alternative factor specifications, security characteristics, and the cross-section of expected returns, *Journal of Financial Economics* 49, 345-373.

Cao, Charles, Eric Ghysels and Frank Hatheway, 2000, Price discovery without trading: evidence from the Nasdaq pre-opening, *Journal of Finance,*55, 1339-1365.

Chan, K., P. Chung, and H. Johnson, 1993, Why options prices lag stock prices: A trading-based explanation, *Journal of Finance* 48, 1957-68.

Chan, Louis, and Josef Lakonishok, 1993, Institutional trades and intraday stock price behavior, *Journal of Financial Economics* 33, 173-199.

Choi, J.Y., Dan Salandro, and Kuldeep Shastri, 1988, On the estimation of bid-ask spreads: Theory and evidence, *Journal of Financial and Quantitative Analysis* 23, 219-230.

Christie, William G., and Paul H. Schultz, 1994, Why do NASDAQ market makers avoid odd-eighth quotes? *Journal of Finance* 49, 1813-1840.

Cohen, Kalman, Steven Maier, Robert Schwartz, and David Whitcomb, 1981, Transaction costs, order placement strategy, and the existence of the bid-ask spread, *Journal of Political Economy* 89, 287-305.

Cooper, I.A. and E. Kaplanis, 1994, What explains the home bias in portfolio investment? *Review of Financial Studies* 7, 45-60.

Copeland, Thomas C., and Daniel Galai, 1983, Information effects of the bid-ask spread, *Journal of Finance* 38, 1457-1469.

Demsetz, Harold, 1968, The cost of transacting, *Quarterly Journal of Economics* 82, 33-53.

Domowitz, Ian, 1993, Automating the Price Discovery Process: Some International Comparisons and Regulatory Implications, *Journal of Financial Services Research* 6, 305-326.

Domowitz, Ian, Glen, Jack and Ananth Madhavan, 1998, International cross-listing and order flow migration: evidence from an emerging market, *Journal of Finance* 53, 2001-2027.

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69-90.

Easley, D., M. O'Hara, and P.S. Srinivas, 1998, Option volume and stock prices: Evidence on where informed traders trade, *Journal of Finance* 53, 431-66.

Easley, David, Nicholas Kiefer, Maureen O'Hara and Joseph Paperman, 1996, Liquidity, information, and infrequently traded stocks, *Journal of Finance* 51, 1405-1436.

Elton, Edwin and Clifton Green,, 1998, Tax and liquidity effects in pricing government bonds, *Journal of Finance* 53, 1533-1562.

Fisher, Lawrence, 1959, Determinants of risk premiums on corporate bonds, *Journal of Political Economy* 68, 217-237.

Forster, Margaret and Thomas George, 1992, Anonymity in securities markets, *Journal of Financial Intermediation* 2, 168-206.

French, Kenneth and Richard Roll, 1986, Stock return variances: the arrival of information and the reaction of traders, *Journal of Financial Economics* 17: 5-26.

Garbade, Kenneth D., and Silber, William L, 1979, Structural organization of secondary markets: clearing frequency, dealer activity and liquidity risk. *Journal of Finance* 34(3) (June), 577-593.

Garman, Mark, 1976, Market microstructure, *Journal of Financial Economics* 3, 257-275.

George, Thomas J., Gautam Kaul, and M. Nimalendran, 1991, Estimation of the bid-ask spreads and its components: A new approach, *Review of Financial Studies* 4, 623-656.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable? *Journal of Finance* 49, 1127-1161.

Glosten, Lawrence R., and Lawrence E. Harris, 1988, Estimating the components of the bid-ask spread, *Journal of Financial Economics* 21, 123-142.

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71-100.

Grant, Dwight and Robert E. Whaley, 1978, Transaction costs on government bonds: a respecification, *Journal of Business* 51,1, 57-64.

Hansch, Oliver, Narayan Y. Naik, and S. Viswanathan, 1998, Do inventories matter in dealership markets: Evidence from the London Stock Exchange, *Journal of Finance* 53, 1623-1656.

Harris, Lawrence, 1986, A Transactions Data Study of Weekly and Intradaily Patterns in Stock Prices, *Journal of Financial Economics* v. 16, 99-117.

Harris, Lawrence, 1989, *A Day-end Transaction Price Anomaly, Journal of Financial and Quantitative Analysis* v.24, 29-45.

Harris, Lawrence, 1991, Stock price clustering and discreteness, *Review of Financial Studies* 4, 389-415.

Harris, Lawrence, 1994, Minimum price variations, discrete bid-ask spreads and quotation sizes, *Review of Financial Studies* 7, 149-178.

Hasbrouck, Joel, 1988, Trades, quotes, inventories, and information, *Journal of Financial Economics* 22, 229-252.

Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *Journal of Finance* 46, 179-207.

Hausman, Jerry, Andrew W. Lo, and A. Craig MacKinlay, 1992, An ordered probit analysis of transaction stock prices, *Journal of Financial Economics* 31, 319-330.

Ho, Thomas S.Y.; Schwartz, Robert A. and Whitcomb, David K. 1985. The trading decision and market clearing under transaction price uncertainty. *Journal of Finance* 40(1) (March) 21-42.

Ho, Thomas, and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 47-73.

Ho, Thomas, and Hans R. Stoll, 1983, The dynamics of dealer markets under competition, *Journal of Finance* 38, 1053-1074.

Holthausen, Robert W.; Robert W. Leftwich, and David Mayers, 1987, The effect of large block transactions on security prices: a cross-sectional analysis, *Journal of Financial Economics* 19, 237-267.

Hong, Gwangheon and Arthur Warga, 2000, An empirical study of bond market transactions, *Financial Analysts Journal* 56, 2 (March/April) 32-46.

Huang, Roger D. and Ronald Masulis, 1999, FX spreads and dealer competition across the 24-hour trading day, *Review of Financial Studies* 12(1), 61-93.

Huang, Roger D., and Hans R. Stoll, 1994, Market microstructure and stock return predictions, *Review of Financial Studies* 7, 179-213.

Huang, Roger D., and Hans R. Stoll, 1996, Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE, *Journal of Financial Economics* 41, 313-357.

Huang, Roger D., and Hans R. Stoll, 1997, The components of the bid-ask spread: A general approach, *Review of Financial Studies* 10, 995-1034.

Jegadeesh, Narasimhan, 1993, Treasury auction bids and the Salomon squeeze. *Journal of Finance* 48 (4) (September), 1403-1419.

Kang, Jun-Koo and René Stulz, 1997, Why is there a home bias? an analysis of foreign portfolio equity ownership in Japan, *Journal of Financial Economics* 46, 3-28.

Keim, Donald, and Ananth Madhavan, 1997, Transaction costs and investment performance: An inter-exchange analysis of institutional equity trades, *Journal of Financial Economics* 46, 265-292.

Kraus, Alan, and Hans R. Stoll, 1972a, Parallel trading by institutional investors, *Journal of Financial and Quantitative Analysis,*7, 2107-2138.

Kraus, Alan, and Hans R. Stoll, 1972b, Price impacts of block trading on the New York Stock Exchange, *Journal of Finance* 27, 569-588.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315-1335.

Lakonishok, Josef, Shleifer, Andrei and Robert Vishny, 1992, The impact of institutional trading on stock prices, *Journal of Financial Economics* 32(1), 23-43.

Laux, Paul, 1995, Dealer market structure, outside competition, and the bid-ask spread, *Journal of Economic Dynamics and Control*, 19, 683-710

Lee, Charles M.C., 1993, Market integration and price execution for NYSE-listed securities, *Journal of Finance* 48, 1009-1038.

Lee, Charles M.C., and Mark J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733-746.

Lee, C.M.C., M.J. Ready, and P.J. Sequin, 1994, Volume, volatilty, and NYSE trading halts. *Journal of Finance* 49, 183-214.

Lee, Ruben, 2002, The future of securities exchanges, in Litan, Robert and Richard Herring (eds.) Brookings-Wharton Papers on Financial Services 2002. Washington, D.C.: Brookings Institution Press.

Lin, Ji-Chai, Gary Sanger, and G. Geoffrey Booth, 1995, Trade size and components of the bid-ask spread, *Review of Financial Studies* 8, 1153-1183.

Lippman, Steven, and John J. McCall, 1986, An operational measure of liquidity, *The American Economic Review* 76, 43-55.

Lyons, Richard, 1995, Test of microstructural hypotheses in the foreign exchange market, *Journal of Financial Economics*, 39 (October), 321-351.

Madhavan, Ananth, 1992, Trading mechanisms in securities markets, *Journal of Finance* 47, 607-641.

Madhavan, Ananth. 1995. Consolidation, fragmentation and the disclosure of trading information, *The Review of Financial Studies* 8, 579-603.

Madhavan, Ananth, 2000, Market microstructure: a survey, *Journal of Financial Markets*, 3, 3 (August) 205-258.

Madhavan, Ananth, and Seymour Smidt, 1991, A Bayesian model of intraday specialist pricing, *Journal of Financial Economics* 30, 99-134.

Madhavan, Ananth, and Seymour Smidt, 1993, An analysis of changes in specialist inventories and quotations, *Journal of Finance* 48, 1595-1628.

Madhavan, Anath, Matthew Richardson, and Mark Roomans, 1997, Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* 10, 1035-1064.

Madhavan, Anath and Venkatesh Panchapagesan, 2000, Price discovery in auction markets: a look inside the black box, *Review of Financial Studies* 13, 627-658.

Manaster, Steven, and Steve C. Mann, 1996, Life in the pits: Competitive market making and inventory control, *Review of Financial Studies* 9, 953-975.

O'Hara, Maureen, 1995, *Market Microstructure Theory*, Cambridge MA: Blackwell.

Pagano, Marco and Ailsa Roell, 1992, Auction and dealership markets: what is the difference?, *European Economic Review* 36, 613-623.

Ready, Mark, 1999, The specialist's discretion: Stopped orders and price improvement, *Review of Financial Studies* 12, 1075-1112.

Reiss, P.C., and I.M. Werner, 1998, Does risk sharing motivate interdealer trading?, *The Journal of Finance* 53, 1657-1703.

Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* 39, 1127-1139.

Scholes, Myron, 1972, The market for securities: Substitution versus price pressure and the effects of information on share price, *Journal of Business* 45, 179-211.

Schultz, Paul, 2001, Corporate bond trading costs: a peek behind the curtain, *Journal of Finance* 56, 677-698.

Sofianos, George and Ingrid Werner, 2000, The trades of NYSE floor brokers, *The Journal of Financial Markets* 3, 139-176.

Stephan, J. and R.E. Whaley, 1990, Intraday price changes and trading volume relations in the stock and stock option markets, *Journal of Finance* 45, 191-220.

Stoll, Hans R., 1978a, The supply of dealer services in securities markets, *Journal of Finance* 33, 1133-1151.

Stoll, Hans R., 1978b, The pricing of security dealer services: An empirical study of NASDAQ stocks, *Journal of Finance* 33, 1153-1172.

Stoll, Hans R. 1979, Regulation of securities markets: an examination of the effects of increased competition," *Monograph Series in Finance and Economics*, 1979-2, New York University, Graduate School of Business, 82 pages.

Stoll, Hans R., 1989, Inferring the components of the bid-ask spread: Theory and empirical tests, *Journal of Finance* 44, 115-134.

Stoll, Hans R., 2000, Friction, *Journal of Finance* 55, 1479 – 1514.

Stoll, Hans R., and Robert E. Whaley, 1983, Transaction costs and the small firm effect, *Journal of Financial Economics* 12, 57-79.

Stoll, Hans R., and Robert E. Whaley, 1990, Stock market structure and volatility, *The Review of Financial Studies* 3, 37-71.

Telsar, L. and I. M. Werner, 1995, Home bias and high turnover, *Journal of International Money and Finance* 14, 467-493.

Tinic, Seha M., 1972, The economics of liquidity services, *Quarterly Journal of Economics* 86, 79-93.

Tinic, Seha M. and Richard R. West, 1974, Marketability of common stocks in Canada and the USA: A comparison of agent vs. dealer dominated markets, *Journal of Finance* 29, 729-746.

U. S. SEC, *Institutional Investor Study*, 1971, Report of the Securities and Exchange Commission, 92nd Congress, 1st Session, House Document No. 92-64 (March 12), Washington: GPO.

U.S. Securities and Exchange Commission, 2001, Report on the comparison of order executions across equity market structures, January, 2001.

Vijh, Anand, 1990, Liquidity of the CBOE equity options, *Journal of Finance* 45, 1157-1179.

Wahal, Sunil, 1997, Entry, exit, market makers and the bid-ask spread," *Review of Financial Studies* 10, 871-901.

Wermers, Russ, 1999, Mutual fund herding and the impact on stock prices, *Journal of Finance* 54(2), 581-622.

Wood, Robert L., Thomas H. McInish, and J. Keith Ord, 1985, An investigation of transactions data for NYSE stocks, *Journal of Finance* 40, 723-739.

Table 1**. Spread measures by market value decile, 1706 NYSE stocks, December 1 1997 – February 28, 1998.-** In cents per share. The quoted half-spread is half the difference between the ask and the bid, averaged over the day. The effective half-spread is the absolute value of the trade price less the quote midpoint averaged over the day. The traded half-spread is half the difference between the average price of trades on the ask side less the average price of trades at the bid side. In calculating the daily average prices, trade prices are weighted by shares traded. The stock price is the closing price. The values in the table are averages over 61 days and over the stocks in each category. Measures of statistical significance are not shown. However, all spread measures are significantly different from zero with every t-ratio exceeding 10.

	Market value decile										
	Smallest	2	3	4	5	6	7	8	9	Largest	Overall
Quoted half-spread	8.28	8.56	8.63	8.27	8.55	7.79	7.30	7.90	6.91	6.49	7.87
Effective half-spread	6.09	6.07	6.11	5.79	6.06	5.49	5.09	5.70	4.87	4.57	5.58
Traded half-spread	3.88	3.77	3.83	3.60	3.71	3.42	3.54	3.89	3.73	4.05	3.74
Roll half-spread	4.49	3.68	3.32	3.33	3.28	3.11	3.08	4.17	3.85	5.18	3.81
Stock price (dollars)	9.33	15.69	22.68	25.20	30.34	32.58	35.58	44.97	50.73	64.45	33.15

Source: Stoll (2000).

Table 2. **Cross section regression of the average proportional half-spread as a function of average stock characteristics in the period, December 1 1997 to February 28, 1998.** Coefficients are in the first line, and t-values are below. LogV is the natural log of the average daily dollar volume. σ^2 is the daily return variance for the prior year. Log MV is the log of the stock's market value at the end of November 1997. Log P is the log of the average closing stock price. Log N is the log of the average number of trades per day. Avg|I| is the average daily percentage imbalance between the volume at the ask and at the bidThe dependent mean and all coefficients except that on σ^2 are multiplied by 100. There are 1706 observations.

	Dep Mean	Intercept	LogV	σ^2	LogMV	LogP	LogN	Avg|I|	Adj R2
S/P	0.389	1.9401	-0.1360	1.5757	0.0400	-0.2126	0.0880	0.0049	0.7974
		21.77	-12.08	18.00	5.75	-18.64	5.45	4.88	

Source: Stoll (2000).

Table 3. **Comparison of execution costs in Nasdaq and NYSE for a matched sample of 175 stocks, based on all transactions in 1991.** In cents. The quoted half-spread is half the difference between the quoted ask and quoted bid. The effective half-spread is the absolute difference between the traded price and the quote midpoint at the time of the trade. The realized half-spread is the five minute price change after a trade at the bid or the negative of the five minute price change after a trade at the ask. The Roll half-spread is the square root of the negative of the mean serial covariance of price changes.

Execution measure	Nasdaq	NYSE
Quoted half-spread	24.6	12.9
Effective half-spread	18.7	7.9
Realized half spread (5 minutes)		
Trades at bid	15.3	2.7
Trades at ask	13.6	0.8
Roll half-spread	18.3	3.4

Source: Huang and Stoll (1996).