

# Wharton

---

Financial  
Institutions  
Center

*The Immediacy Implications of  
Exchange Organization*

by  
James T. Moser

02-11

The Wharton School  
**University of Pennsylvania**





## The Wharton Financial Institutions Center

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

  
Franklin Allen  
Co-Director

  
Richard J. Herring  
Co-Director

*The Working Paper Series is made possible by a generous  
grant from the Alfred P. Sloan Foundation*

# The Immediacy Implications of Exchange Organization

by

James T. Moser

Research Department  
Federal Reserve Bank of Chicago  
230 S. LaSalle Street  
Chicago, IL 60604

[jmoser@frbchi.org](mailto:jmoser@frbchi.org)

January 31, 2001

## **ABSTRACT**

The paper introduces a connection between the needs of exchanges to respond to the immediacy needs of their clientele and the need to manage the credit risks faced by exchange members. Queueing theory is used to represent the opportunity loss suffered by brokers engaging in multiple activities: order-flow origination and its intermediation. The role of market-making locals is depicted as enabling specialization. Brokers focus on originating order flow and locals on fulfilling intermediation needs. The capacity to specialize is constrained by the availability of creditworthy members acting as locals. This results in a tension between pursuit of immediacy and managing inter-member credit exposure. Two exchange rules, tick size and price limits, are evaluated for their effects in resolving this tension.

This research benefits from the comments of Ray DeGennaro, Mark Flannery, Steve Kane, Tom Lindley, Jay Marchand, Pat Parkinson, Asani Sarkar, Lester Telser, Rich Tsuhara and participants of the Brookings-Wharton Financial Services Conference (January, 2002). Errors remaining in this draft are mine. The views of the paper do not reflect the official positions of the Federal Reserve.

## **I. Introduction**

Moser (2000) describes the development of modern futures clearinghouses as the culmination of a series of adaptations to credit risk problems. Baer, France and Moser (2001) extend this reasoning to show how contract margin requirements reflect the interests of exchange members who jointly minimize their credit risk exposures and their margin-carry costs. Common to these papers is their application of a theory of efficient contract design to futures contracts. A contractual perspective on exchange membership is also useful. Viewed this way, exchange rules become terms to contracts of membership. In turn, those rules become subject to the same efficiency concerns as any other contract-design problem. This thought exercise gives insight into the economics that underlay exchange organizations. Just as members continually re-contract to control price risks arising from their extant financial positions, they also adjust exchange rules affecting their membership standings.

This paper models linkages between exchange decisions affecting contract performance and meeting customer demand for immediacy. Immediacy is the time required for filling customer orders and, in this sense, the term describes the performance of transaction service providers. The duration of order-fill processing creates exposure to price changes, an exposure that increases as order-fill duration lengthens. This outcome is attributable to both information arrival and to market liquidity. More straightforward is information arrival: Informed traders want their orders filled before prices adjust to their information. Less straightforward is the added exposure to the price implications

of liquidity. Liquidity is a conditional expectation of the association between the price response of a transaction and its size. As large transactions are more likely to move prices, liquidity descriptions are usually made in terms of the largest transaction having no expected price impact. Because long-duration order fills increase exposure to these price effects, poor liquidity elevates the importance of immediacy.

Immediacy needs are common to all exchanges. Securities markets are organized to facilitate brokered transactions with a combination of human and computer resources. Specialists holding limit order books provide a nexus for price discovery. Computerized order-handling meets most immediacy needs by using algorithms to match the buy and sell sides of most at-market orders. Open-outcry futures markets are organized as continuous auction markets. Immediacy is provided by locals tracking buy and sell interests on the floor; the transactions of locals serve to intermediate externally originating order flows.

Given the problems arising from a lack of immediacy, exchange members have interests in taking steps that reduce the problem. Increasing the number of members improves the number of opportunities to find a counter party within a time span. Selecting from a homogenous population of potential members, the choice is straightforward: add members until the immediacy problem goes away. However, potential members are more likely to be heterogeneous, particularly so in their credit dimensions. Scarcity of strong credits among a population of potential members constrains immediacy improvements. This paper uses a queueing theory representation to quantify costs implied by inadequate

immediacy and develops a trade off between these costs and the costs incurred when weak credits are accepted as members.

To see its relevance, consider how membership size affects an exchange member's interests. We readily see the benefit when our grocer opens another checkout line and decreases the portion of a Saturday spent standing in line. Likewise, adding an exchange member can decrease time spent locating contract counterparties. This is to say that increasing the number of members improves transaction immediacy at a given price. From this, it follows that the cost for obtaining immediacy declines as membership size increases. From the perspective of individual members, every other member is a potential service channel and—like grocery check out lanes—more service channels are preferred to less.

On the other hand, adding members has risk management consequences. Baer, France and Moser (2001) show that monitoring members for their nonperformance prospects can substitute for collateralizing against losses. However, they find no evidence of differential collateral assessments as would be indicative of substantive reliance on monitoring. Collateral appears to be the primary risk-management tool and collateral requirements are the same for most members. This being the case, an increase in the riskiness of new members increases the collateral required from all members.

The literature on financial intermediation has not previously drawn from queueing theory to address financial structural issues. Indeed, use of the theory is infrequent in the general economic literature. Naor (1969) includes queue-

residence time as a component of all-in product cost. De Vany (1976) introduces queueing theory to the industrial organization literature by incorporating wait time into monopoly pricing problems. In their analysis of the trucking industry, De Vany and Saving (1977) extend wait time costs to pricing in competitive markets. The empirical study of Frech and Lee (1987) gauges the inefficient allocation of gasoline caused by use of wait time as a non-price rationing mechanism. Davidson (1988) shows how firms can use wait time preferences to segment service markets.

The next section illustrates concepts of the paper with a numerical illustration for an exchange comprised entirely of principals. Section III models an exchange whose order flow originates externally. Specialization arises as brokers work with external parties to develop order flow. Immediacy providers absorb order flow based on their abilities to track buy and sell interests among the brokers. Section IV relates exchange rules to efforts to improve the terms of the trade off between immediacy demands and exposure to risk. Price tick rules affect the share of revenue obtained by locals, adjusting these rules can affect local participation in a contract. Price limit rules limit the level of liquid resources required of locals. Section V develops some perspective on the policy implications of the model and summarizes the paper.

## **II. Exchange Membership Comprised of Principals**

A simulation exercise provides an intuitive basis for this queueing representation. A population of potential exchange members is constructed and

ordered by the expected losses incurred by surviving members when any single member fails. For this exercise, members trade for their own account only, that is, they trade as principals. Assuming the clearing organization requires full protection from expected losses, the required collateral level is the amount of loss exposure implied by the weakest credit permitted to join. Calibrating loss amounts to an S&P transition matrix of bond ratings adds some realism to the exercise. For each of the possible memberships of size  $N$ , the collateral required of all members is the expected loss on failure of the weakest credit. The opportunity cost is the alternative return that members can earn by investing their collateral elsewhere. For purposes of the simulation, this is 5% scaled to a contract value of 1000.

As suggested by the above discussion, adding members increases the number of service channels and decreases expected wait time. Using standard queueing results<sup>1</sup>, I calculate expected wait times for each possible membership size. Assuming wait time is mutually exclusive of other productive activity, the cost of time in the queue accrues at 5% per period. I normalize costs to the servicing cost for one contract servicing period. Thus, the all-in cost for one contract is the 5% incurred during the servicing interval plus a charge for time spent waiting. If the time spent waiting is 10% of the time spent processing, then the additional charge is  $\frac{1}{2}\%$  ( $= 5\% \times 1/10$ ).

Table 1 gives the calculated cost schedules for collateral and immediacy. Collateral costs rise gradually from zero to 0.0367 per contract. The

---

<sup>1</sup> See Moser (2002).



steepness arises from the sharp increases in average default losses as bond ratings decline. In contrast, adding new members drives immediacy costs to zero very quickly. For this illustration, combined costs—by construction these are average costs—reach their minimum at five members. Beyond five members combined costs rise—the steep rise in collateral-holding costs dominating a less rapid decline in immediacy costs.

**Table 1**

Number of Members	Per Member Costs		
	of Collateral	For Immediacy	Combined
2	0.0033	0.2963	0.2996
3	0.0038	0.0164	0.0202
4	0.0043	0.0032	0.0075
5	0.0050	0.0007	0.0057
6	0.0060	0.0002	0.0062
7	0.0075	0.0000	0.0075
8	0.0100	0.0000	0.0100
9	0.0150	0.0000	0.0150
10	0.0300	0.0000	0.0300
11	0.0360	0.0000	0.0360
12	0.0367	0.0000	0.0367

The next section moves away from this specific parameterization and moves toward developing economic intuition for this perspective on a futures exchange.

### **III. Specialization Within an Exchange Membership**

Memberships do not specify the activities of individual members, but members do specialize. Though the term “market maker” connotes a single market for all buy and sell activity, in reality “market makers” construe their markets much more narrowly. One market maker may be quoting a June-

September spread market while another quotes a market for the September contract only.

This role differentiation also applies to the activities required for servicing orders arriving from nonmembers. Brokers and Futures Commission Merchants (FCMs) specialize in bringing buy and sell orders to the exchange, locals specialize in supplying immediacy. This section expands on this differentiation and models the provision of immediacy services.

#### The Immediacy Provisioning Activity of “Locals”

Brokers focus on bringing order flow to market. This activity precludes time spent keeping track of extant buy and sell interests. As a result, brokers are not well equipped to immediately match a new buy or sell order to another broker seeking to sell or buy. Unlike brokers, locals do not bring order flow to market. Locals fill the need for immediacy by keeping themselves aware of the current buy or sell interests of the brokers. This role separation constitutes a specialization of skills. Brokers specialize in developing and maintaining the external associations needed to bring order flow onto the exchange and locals specialize in matching buy and sell orders.

Locals accomplish their roles by transacting with brokers to take pieces of broker-originated customer orders. Upon its being parceled out to locals, broker servicing of a customer order is complete. Locals, in turn, then transact to reverse their positions. Those transactions ultimately match to customer orders on the other side. For example, broker A has been negotiating with a customer for a large sell order. In the interim, the buy-sell interests of other brokers could

change, so that the task of locating buyers can delay order execution. Locals tracking the buy interests of floor brokers will buy portions of the order originated by broker A then sell the position to brokers presently looking to fill orders for their customers.

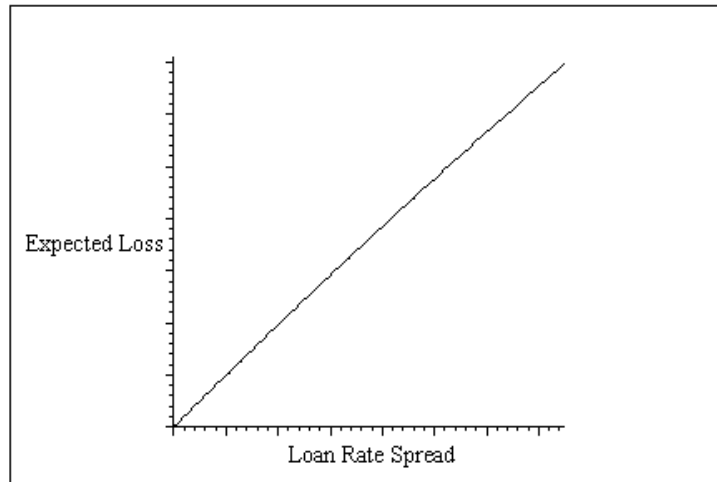
The specialization of the activities of locals and brokers requires revising the representation of the exchange provided in the previous section. Orders awaiting service are those held by locals; that is, the waiting area of the queue consists of the positions held by locals. This implies a capacity limitation on the number of orders queued at any instant. The combined financial capacity of locals to carry positions defines this limit with new orders being rejected on reaching this limit.

### The Cost of Adding Liquidity Providers

Loan spreads of the form  $r - r_f$  compensate for expected loss plus an add-on compensating for the risk that actual loss exceeds the expectations of risk-averse lenders. With an appropriate model, one can parse spread components to extract the loss expectation. The point of this subsection is to demonstrate how member differences in expected losses affect their opportunity costs. Thus, assuming risk neutrality is sufficient for present purposes and expected loss is represented as  $e^{(r_0 - r_f)} - 1$  where  $r_0$  is the rate for loans extended to member 0 and  $r_f$  is the default-free rate. Save for their different default prospects, the loan terms priced in the two rates are matched in all relevant respects and both are continuously compounded.

Observing the loan spreads available to potential members provides the exchange with useful information on their creditworthiness. Computing expected losses as above and ranking these lowest to highest (see Figure 1) provides the exchange with information relevant to its future collateral requirements.

**Figure 1**



**Ranked Expected Member Losses**

On reviewing applicant  $i$ , member 0 favors the application when the following holds:  $e^{(r_i-r_f)} \leq e^{(r_0-r_f)}$ . This rule holds because adding the new member implies no increase in the costs of risk management. When  $e^{(r_i-r_f)} > e^{(r_0-r_f)}$  the new member poses a risk increase that must be managed and the membership incur costs associated with managerial effort.<sup>2</sup> Retaining the presumed risk neutrality, margin collected from the new member covers the loss expected from its contract nonperformance. Comparing a prospective new member to member 0, the

<sup>2</sup> Use of a potential member's existing borrowing rate implies that exchange access to information about a potential member is no better than that known by the member's lenders. Neither does the exchange have a comparative advantage in using the information it does have.

amount of extra margin to be collected is  $e^{(r_i-r_f)}-e^{(r_o-r_f)}$ . When margin assessments are member specific, then collecting margin sufficient to cover the expected loss owing to that member's nonperformance does not increase costs for existing members. It follows too that when members cover risk-management costs introduced by their membership, prospects improve that existing members are more likely to favor their admission.

However, exchanges do not differentiate their margin assessments. Instead, margin requirements are uniform implying that a membership may realize higher costs when admitting new members.<sup>3</sup> Member 0 calculates the cost added by admitting an additional member from three items: the increase in required margin, the rate of return that member 0 can earn on this amount, and the number of contracts on which the added amount will apply.<sup>4</sup> In a one-member-one-vote organization, members evaluate their benefit from adding new members. Decisions based solely on extra margin cost and immediacy improvements will admit new members up to the point where at least  $N/2+1$  voting members expect positive net benefits.

---

<sup>3</sup> Though exchanges generally set uniform margins, clearing members can and do assess higher margins for the accounts they clear. This does obtain greater differentiation than implied by exchange rules. The conclusions of this paper go through provided margin assessments are less than perfectly elastic with respect to the nonperformance risks of individual members. I am grateful to Pat Parkinson for pointing out this institutional detail.

<sup>4</sup> By reducing the number of open contracts, member 0 can reduce the cost implied by the added amount and the rate of return available on that amount. However, this entails a reduction in benefit that is otherwise obtained by carrying those positions.

The preceding characterization demonstrates a relationship between nonperformance costs and the number of memberships. Exchanges add brokers to bring order flow to the exchange, they add locals to improve the immediacy of order fills. Order flow and immediacy are linked through the extent to which immediacy permits brokers to engage in their specialties. Because the value added from order-flow origination is clear, the remainder of the paper focuses on the immediacy implications of adding locals.

### The Value Added by Liquidity Providers

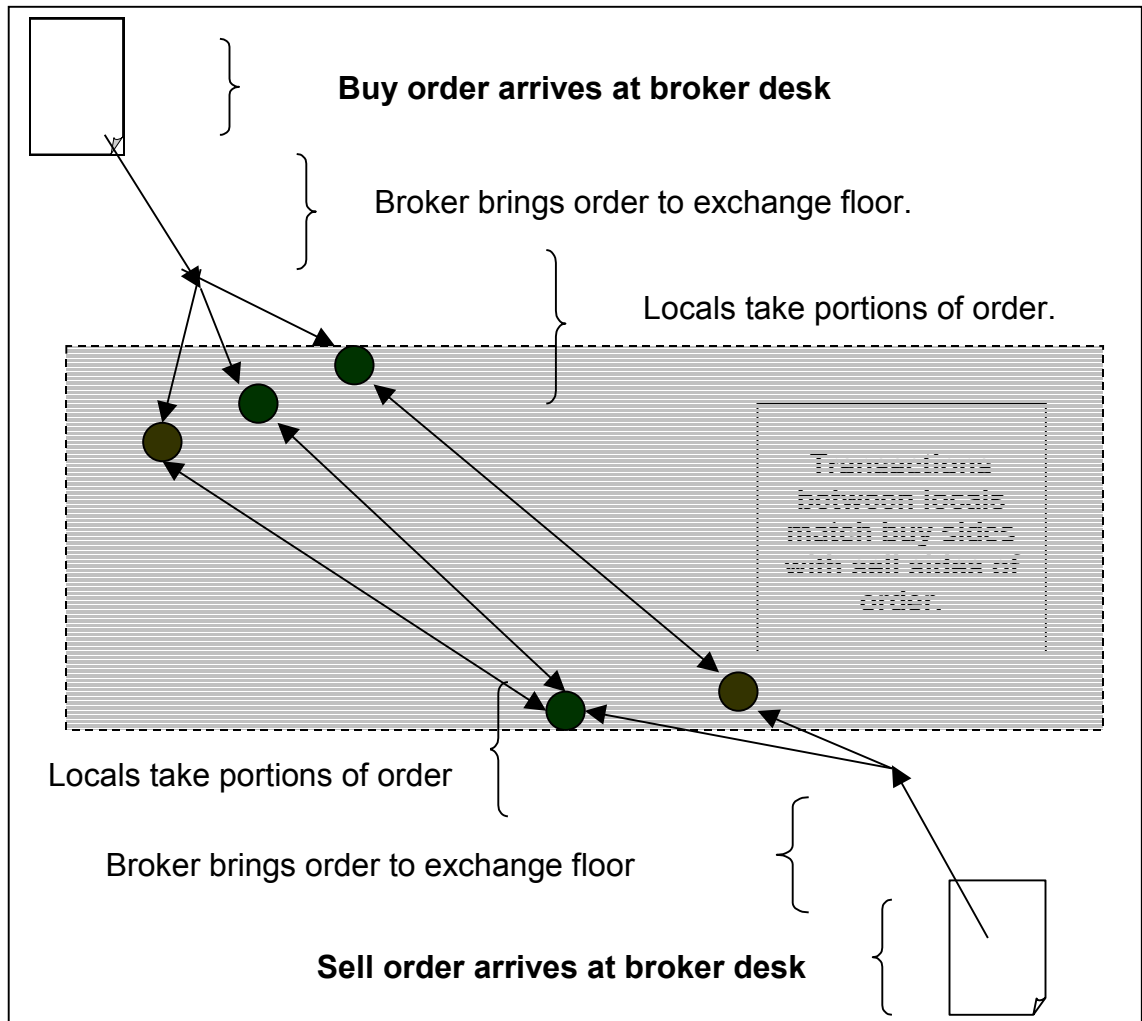
Locals add value by absorbing short-term, order-flow imbalances. A broker bringing a large order to the floor best serves client interests when the order is executed at an average price no less favorable than the market price when the order was given. Simultaneous arrival of identically sized and offsetting orders—a buy for every sell—is unlikely. More likely, the requisite order matches are dispersed amongst several brokers. “Working the order” is time consuming; that is, time spent locating selling brokers and negotiating prices detracts from the broker’s efforts to bring additional order flow into the exchange. Locals, specializing in tracking existing buy and sell interests on the floor, take positions expecting to trade out of them.<sup>5</sup> The latter trade can be with a selling broker or another local.<sup>6</sup> Figure 2 provides a schematic detailing the order flow being modeled.

---

<sup>5</sup> Locals carry positions for very short intervals. Silber (1984) calculates average holding periods for a sample of locals. He reports the average term of their positions is less than two minutes.

<sup>6</sup> Trades with other locals are more frequent. Curran (2002) shows that during a 1997 sample period broker-to-broker transactions in the S&P 500 contract

Figure 2



For simplicity, I assume that one or more locals immediately take up orders introduced by brokers. In practice, these transactions are less immediate. However, anecdotal evidence suggests this time is generally much smaller than the time spent by locals working their pieces of the order. The assumption

---

represented 10% of contract volume. All other transactions were broker-to-local (65%), or local-to-local (25%).

permits a simplification: I can represent order flow taken by locals as queued for matches with later-arriving orders.<sup>7</sup>

Initially, I also assume that locals take, at most, one contract. In queueing theory terms, this implies that a local represents one potential queue space. The assumption avoids two complications. First, it enables the analysis to focus on the marginal effect from adding queue spaces rather than the marginal effect of adding locals with each taking more than one order. Second, the assumption sidesteps comparisons of contract-nonperformance prospects for locals taking multiple contracts.<sup>8</sup> On establishing a tradeoff between immediacy and credit risk, the succeeding section then considers how exchange rules can affect the capacity of locals to carry more positions.

Moser (2002) provides performance measures for a queueing model having  $c$  service channels and  $N-c$  queueing spaces. The model has two stochastic elements: customer inter-arrival time and servicing times, both assumed to be Poisson distributed. I denote expected inter-arrival times as  $\lambda$  and expected service times denoted as  $\mu$ . With these parameters and assuming a steady state, an expectation for queue length, denoted  $L_q$ , can be derived. The next three figures convey intuition for the effects from changing service channels (broker) and number of queue space (locals).

---

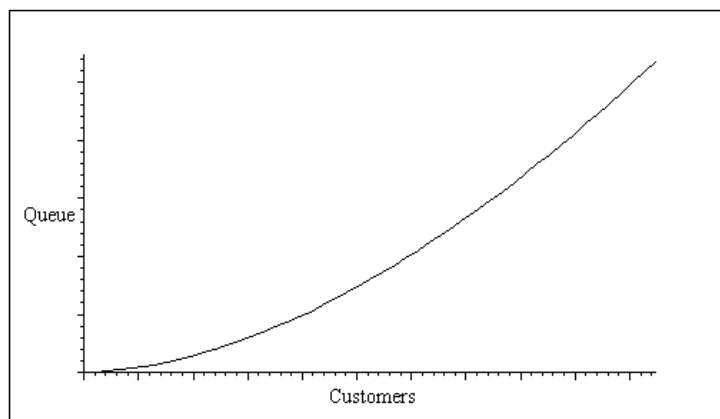
<sup>7</sup> Queueing theory does offer ways to deal with a series of queues. I find no gains in insight from adding this complexity.

<sup>8</sup> For example, a local having weak credit carrying one contract can pose less credit exposure than posed by a local having stronger credit but carrying 1,000 contracts.



Figure 3 illustrates the effect on queue length from adding customers with a fixed number of servicing channels. Queue capacity is defined as the maximum number of customers that can be in the system minus the number in service channels. The curve's flatness near the origin occurs because expected queue length is zero when the number of service channels exceeds the number of customers. Beyond that point, expected queue length rises as the number of customers in the system rises. For an exchange, the figure demonstrates that for a fixed number of brokers (service channels), the need for locals to absorb arriving orders increases as the number of customers increases.

**Figure 3**

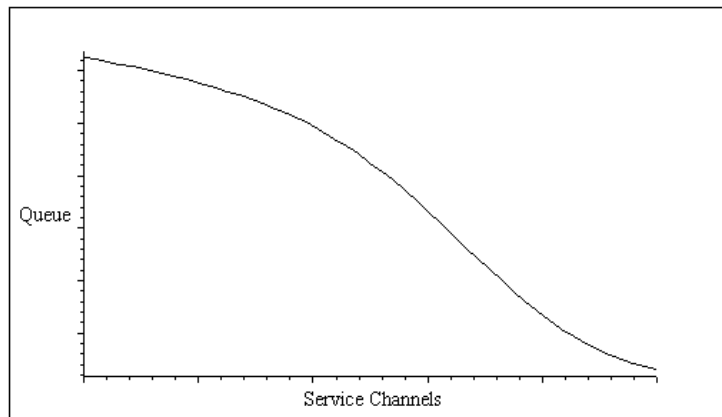


Effect of Adding Customers on Queue Length

Figure 4 illustrates the effect on queue size from increasing the number of service channels for a fixed number of customers. The initial effect on queue size from increasing the number of service channels is, at first, relatively small. As increasing the number of service channels improves the odds that arriving customers are immediately processed, expected queue size declines more rapidly. As the probability increases that a service channel will be open, queue

length declines less rapidly. As before, a service channel is interpreted as a broker. Hence, increasing the number of its active brokers diminishes the need for locals to carry positions.

**Figure 4**



Effect of Adding Channels on Queue Length

The following differential equation combines the effects of increasing queue capacity ( $dN$ , changes in the number of locals) and adding service channels ( $dc$ , changes in the number of brokers):

$$dc \frac{\partial L_q}{\partial c} + dN \frac{\partial L_q}{\partial N} = 0 \quad (1)$$

Solving for  $dN/dc$  obtains the needed increase in locals for a change in the number of brokers to obtain the same expected queue length. I evaluated the expression for various numbers of locals and brokers, finding that the number of locals increases, at increasing rates, in the number of brokers.

## Cost-Minimizing Behavior

The previous two subsections identify quantities for two costly resources. Managing credit risk entails pledging collateral against nonperformance. The cost of this resource is its use in other productive activities. Time, in particular time spent queued, is the second resource and its cost is the value of foregone activities. To equilibrate these resource costs, I re-state expected queue length as expected queue time as follows:

$$W_q = \frac{L_q}{\lambda} \quad (2)$$

The intuition for the transformation is as follows. In equilibrium expected customer arrivals occur every  $\lambda$  time units. During that interval, those previously residing in the queue expect to advance one space. A counter example makes the case for this result. Were it not true, the system cannot be in equilibrium as either expected queue length continues to increase or the system goes to zero. Hence, the expected time cost is the value of lost opportunities during queued intervals of length  $W_q$ .

At the margin, the expected rate of return from foregone opportunities equals the expected rate of return from investment opportunities.<sup>9</sup> Were they not equal, resources would be re-allocated until meeting the equality condition. Hence, the expected wait-time cost for member 0 is  $e^{r_0 W_q} - 1$ . Naturally wait time costs are increasing in  $W_q$ , the rate of increase is  $r_0 e^{r_0 W_q} > 0$ . As shown previously, members can decrease queue length by increasing the number of

---

<sup>9</sup> As before, risk neutrality is assumed.

service channels. It follows that increasing service channels obtains lower wait time costs.

However, scarcity of creditworthiness implies that adding members may require relaxing credit standards and lead to higher costs via greater risk management effort. For simplicity, the exchange manages credit risk entirely through collection of collateral deposits against contracts. Note that this implies that all members post identical collateral amounts. Hence, adding members increases costs for all members. The amount of cost increase incurred when member 0 takes on a new contract is:

$$(e^{r_i - r_f} - e^{r_0 - r_f})(e^{r_0} - 1) \quad (3)$$

that is, the product of the added amount of collateral required from member 0 when member  $i$  is admitted and the opportunity cost paid by member 0 when posting additional collateral. This cost increases in  $r_i$  at the rate  $e^{r_i - r_f + r_0} > 0$ .

Adding members requires comparing the effect of new members on wait-time costs that decrease in membership size and on risk management costs that increase in membership size. The optimal decision for member 0 is when the net cost changes from admitting the new member is zero. This occurs when the following holds,

$$dN \left[ r_0 e^{r_0 W_q} \frac{\partial W_q}{\partial N} \right] + dr_i \left[ e^{r_i - r_f + r_0} \right] = 0 \quad (4)$$

When an existing membership makes this decision each member evaluates their net cost. Proposed new members are admitted when the above condition is satisfied for more than  $N/2$  members.

**Figure 5**

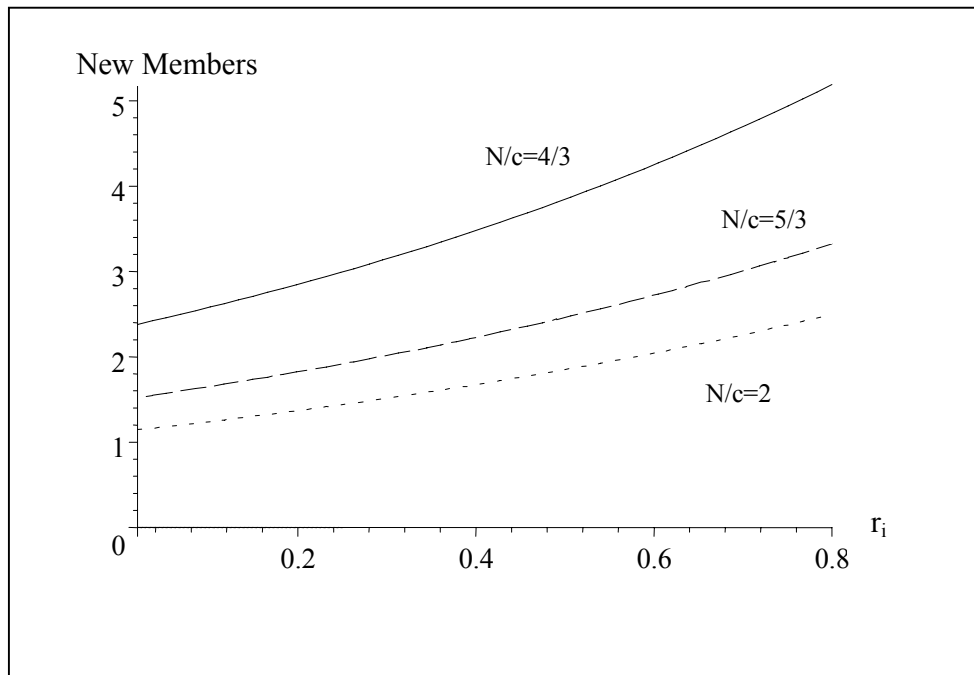


Figure 5 illustrates the margins for these decisions by plotting the  $dN/dr_i$  curves implied by equation 4. The solid curve represents an exchange having four members, three of which are brokers. The dashed curve increases the number of members to five, and the dotted curve increases that number to six. The curves are positively sloped, demonstrating the value placed on improving immediacy especially when the number of locals is small. Existing exchange members are willing to incur substantially higher collateral costs when adding members to improve immediacy. Adding new members at lower levels of credit

risk implies existing members realize improved immediacy with lower costs of risk management.

As the number of existing members ( $N$ ) increases (holding  $c$ , the number of brokers, constant), adding members becomes less attractive. For example, despite the credit risk implied by a new member whose cost of funds is 20%, members value the improvement in immediacy sufficiently to add nearly 3 locals (solid line) when the number of existing members is 4.<sup>10</sup> However, with five existing members (dashed line) less than 2 locals are added. Finally, with six members (dotted line) interest in adding locals declines further still.

In conclusion, when immediacy services are scarce, exchange members are willing to pay substantially higher risk management costs to obtain improvements. As the scarcity of this service declines, their willingness to bear these costs also declines.

#### **IV. Exchange Rules Facilitating Entry of Liquidity Providers**

The previous section illustrates how exchange memberships resolve tensions between their immediacy needs and the effect that fulfilling those needs has on risk management costs. This section explores the separate interests of brokers and locals to obtain greater insight into the members' decision process. A straightforward maximization of the expected profit argument establishes a

---

<sup>10</sup> The model's construction doesn't permit consideration of partial memberships. However, exchange rules permitting "dual trading" do accomplish this. Although subject to certain customer-protection rules, exchanges permit members to trade as both a broker and a local. The Chicago Mercantile Exchange generally prohibits dual trading "in any contract month which is mature and liquid." (CME Rule 552.B) This is to say that the CME permits dual trading where immediacy services are most valued.

relatively strong motivation for brokers to resolve immediacy problems. The case for locals seeking similar ends relies on concessions in the form of exchange rules that favor their market-making activities by increasing revenues and limiting their amounts of required cash capital. These rules are consistent with exchange pursuit of immediacy. For example, the rules might be conjectured as concessions offered by brokers to attract greater participation by locals.

### Broker Interest in Immediacy

Earlier sections use a specialization argument to motivate the interests of brokers for their pursuing immediacy. Allocating their time between two mutually exclusive activities—working with external customers and market making—immediacy provided by locals permits brokers to work within their specialty.

Simple profit maximization strengthens this argument. Recalling the finite queue size developed in the previous section implies, on occasion, an exhaustion of the capacity of locals to accept positions. In these instances, brokers must choose between refusing the orders of current customers or accepting those orders and conducting their own market-making activities. I will assume that in expectations the cost of either is the same. This is tantamount to assuming the expected time to work an order precludes accepting the next order as well as assuming equal value added from either activity. Accepting the second of these assumptions seems reasonable, the first is less reasonable for several reasons. First, brokers working their orders have stepped out of their specialties. It is unlikely that their efforts in a market-making capacity will be equally profitable. Second, a market that has used up its immediacy capacity is

likely to be a fast market. In these instances, time required to work the order is likely to increase while time between order arrivals decreases. Hence, the assumed equality of costs for the brokers' alternatives probably understates the brokers' interests for improved immediacy.

Recalling that  $N$  is defined as system capacity, let  $P_N$  be the probability of  $N$  customers being present; i.e., all service channels and queueing capacity are in use. Having no additional capacity, brokers must reject new orders until queue space becomes available. Queueing theory refers to this as "balking." In such cases, broker profit is nil. During less congested periods, brokers obtain profit denoted  $\pi$ , these times occur with probability  $(1-P_N)$ . Hence, expected broker profits can be stated in queue-capacity terms as follows:

$$E[\pi] = P_N 0 + (1 - P_N)\pi \quad (5)$$

Increasing queue capacity clearly reduces the probability of balking and increases profits, i.e.,  $\frac{\partial E(\pi)}{\partial N} > 0$ . This establishes a motivation for brokers to incur costs in their efforts to resolve immediacy problems. Among these costs can be exchange rules that improve immediacy, most especially rules that improve immediacy during fast markets.

#### Capacity Limits for Locals

Define the local's financial position as  $V$  defined in terms of the local's cash position  $C$  and the market value of her open positions, that is as follows:

$$V = C + FQ_s - FQ_L \quad (6)$$



Where F is the futures price and Q gives the number of open contracts, the subscripts S and L denoting, respectively, short and long positions. Price changes are marked against the local's position. On obtaining sufficient cash to support her positions, then the local's goal is  $\Delta V=0$ . Absent inventory adjustments in response to price changes, then changes in the value of the position are:

$$\Delta C = -\Delta F Q_s + \Delta F Q_L = \Delta F (Q_L - Q_s) \quad (7)$$

Taking expectations and squaring both sides obtains the variance of cash holdings:

$$\sigma_C^2 = \sigma_F^2 (Q_L - Q_s)^2 \quad (8)$$

So the extent of price variability resulting from the local's net position determines the volatility of changes in cash holdings. The next two subsections provide two routes for affecting the local's willingness to increase the supply of immediacy services.

#### Tick Size as a Means of Compensating Locals

Exchanges define tick sizes as the minimum amount of nominal price change. Table 2 illustrates with examples for several well-known contracts.

**Table 2**

Exchange	Contract	Tick Size	Dollar Amount Of Tick Size
CBOT	Treasury Bonds	1/32	31.25 \$
	Soy Beans	1/4	12.50 \$
CME	S&P 500	1/10	25.00 \$
	Eurodollar	1/100	25.00 \$
NYMEX	Crude Oil	1/100	10.00 \$
	Unleaded Gasoline	1/10000	4.20 \$
NYBOT	Orange Juice	1/100	7.50 \$
COMEX	Gold	1/10	10.00 \$
EUREX	Eurobund	1/10000	10 EU

The pricing units used by the underlying cash markets determine tick sizes. This convention, because it eliminates the need to restate futures prices, facilitates futures trading by parties having on-going positions in the underlying market. The product of the pricing interval and the notional value of the futures contract obtain the dollar values implied by each minimum tick size. For example, the S&P 500 pricing interval is 1/100 of an S&P point. The notional value of that contract is 250 times the S&P 500, the product ( $0.01 * 250$ ) is \$2.50. The minimum tick size is 10 pricing intervals for a dollar value of \$25.00.

Consider two possibilities: the contract trades at its pricing interval of 2.50 or at its minimum tick size of 25.00. The bid-offer prices quoted by locals will be multiples of 2.50 or 25.00. At a 2.50 pricing increment, the bid-offer spread can be bid down as low as 2.50. At 25.00, the bid-offer spread can only be bid down to 25.00. In exchange parlance, both are “one-tick markets.” The first implies local compensation is \$2.50 per trade, the second gives compensation of \$25.00 per trade. The effect of choices between these alternatives on the supply of locals is straightforward.

Tick-size contract specifications amount to a form of price-administration having two effects.<sup>11</sup> First, recognize that perceived instances of market failure are an important motivation for administering prices. In the present instance, the minimum tick size assures that the supply of immediacy-providing locals will be higher than may result were the minimum lower. In terms of the value of a local's position, price administration prevents the local from incurring a market-determined bid-offer rate below 25.00. Recalling the earlier point that exchanges do not manage member specialties, the exchange must rely on incentives to adopt needed specialties. This floor on per-contract compensation improves the revenues of locals provided contract volume is sufficiently inelastic to its consequent bid-offer spread.

Second, noting that price variation accumulates over time, minimum price increments increase the time a local has to reverse out of a position. Referring to the above example, define average deviations in equilibrium prices for one-minute holding periods as 2.50. A local can expect no less than 5 minutes to offset a position before equilibrium price changes can be expected to round up (or down) to the next allowable pricing increment of 25.00. Were the minimum price interval to be 2.50, the local has less than 1 minute to do so.<sup>12</sup> Locals seeking to avoid exposure to price changes will prefer contract terms that offer sufficient time to close out their positions

---

<sup>11</sup> For an example, see the Bollen, Smith and Whaley (2001). They study changes in the terms of the CME's S&P 500 contract. Curran (2002) extends their study.

## Price Limits as a Means of Reducing Costs of Carrying Cash Balances

Miller and Orr (1966,1968) derive optimal balances for zero-drift cash accounts when cash holdings have an opportunity cost denoted  $v$  and replenishing account balances costs  $\gamma$  per transaction. The optimal balance  $z^*$  is

$$z^* = \left( \frac{3\gamma\sigma_c^2 t}{4v} \right)^{1/3} \quad (10)$$

where  $\sigma_c^2$  is the variance in dollar terms of net inflows and outflow of cash.

Substituting for  $\sigma_c^2$  with the variance of futures prices (see equation 8) and taking the derivative with respect to  $\sigma_F^2$  gives the effect of a change in price volatility on the position:

$$\frac{\partial z^*}{\partial \sigma_F^2} = \frac{1}{3} \left( \frac{3(Q_L - Q_s)^2 \gamma t}{4v} \right)^{-2/3} > 0 \quad (11)$$

The local's net position is  $(Q_L - Q_s)$ , so from 11 we can conclude that the optimal level of cash balances rises with the volatility of prices, with time, and with cash-balance replenishment cost. The optimal cash-balance level falls as the opportunity cost for cash balances rises.

---

<sup>12</sup> These examples presume tick-size conventions obscure equilibrium prices. Thus, equilibrium prices fall within a range defined as  $\frac{1}{2}$  pricing increment above or below an observed price.

**Figure 6**  
**Optimal Cash Balances and Volatility**

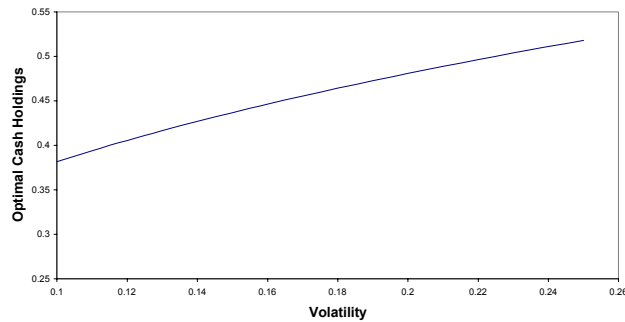


Figure 6 depicts the effect of futures price volatility on optimal cash balances for a single contract over one holding period with a one-period opportunity cost of 5% of the cash replenishment cost. The relation's steepness suggests that price volatility has an important role in determining the activities of the local. Holding constant revenue opportunities per dollar of available cash balances, locals prefer low over high volatility positions. It is appropriate to understand this decision as a long-run choice. For the short run, locals can lessen their needs for cash balances by adjusting their net positions.

Similarly, Telser (1986) relates the short-run adjustment process to the isomorphism between cash and futures contracts introduced. Facing a short fall in cash balances, the local chooses between an inventory adjustment and paying  $\gamma$  to replenish the cash account. Thus, reducing instances requiring replenishing of cash balances lessens the need for locals to adjust inventory.

Price limits establish ceilings on the cash amounts that can be required within a single trading day. To see this, consider a contract whose specifications call for dropping the price limit on the day following a day in which the close price

is the limit price. This, implies that  $\sigma_c^2(t) \leq .5\sigma_c^2(2t)$  where the arguments in parentheses define the intervals for which variances are computed.<sup>13</sup> As demonstrated above, this reduction in cash flow variance reduces the level of cash needed to support the position. The effect is obtained by smoothing cash flow needs across days.

Increasing the period over which a price shock can affect cash balances benefits a local in two ways. First, when the change results from an excessive response to a price shock, a price reversal can be expected. The local benefits by a reduction in cash needed to restore the account. Second, increasing the amount of time available to replenish the account increases opportunities to liquidate other holdings or to re-negotiate loan terms. In both instances, as developed in Moser (1986), the local has the equivalent of a zero-interest loan from the winning side of the contract. Reducing the effective rate for carrying cash balances decreases the cost locals incur when they carry positions during volatile markets. Absent this cost reduction, locals might be expected to exit when prices become volatile having the effect of reducing immediacy as volatility rises. Kuserk and Locke (1996) show that locals are more likely to carry overnight positions when price limits are hit than when limits are not hit. This suggests the value of these implicit loans is sufficient to alter the behavior of locals, they continue to supply immediacy despite the risk they will face overnight exposure to price changes.

---

<sup>13</sup> The inequality is weakened when the probability of a price change reaching the limit is nil.

## V. Policy Implications and Conclusion

Economic models affect policy choices by altering the way that we understand economic choices. Knowledge that a tax distorts choices in predictable ways gives the policy maker a sense of the cost and benefit implied by changes in the tax. It follows that improving the predictability of a response improves the policy maker's effectiveness.

This is particularly important for the regulatory style referred to as "incentive compatible." This regulatory style rewards regulated firms for their compliance with social goals. As well-structured reward systems require full understanding of organizational structure, models that parsimoniously convey structural understanding are more conducive to effective regulatory policy. The key contribution of this paper comes through its illustration of linkage between immediacy levels supplied by exchange membership and the creditworthiness of their membership. Incorporating this perspective into the policy formulation process can improve its effectiveness at minimum by avoiding policy choices that conflict with already-present incentives.

### Policy Implications

The fundamental cause(s) of the Market Break of October 1987 will probably never be understood, in part because the market mechanism itself became part of the problem.<sup>14</sup> What comes through clearly is the dramatic decline in the capacity of market makers during the Break. At the Chicago

---

<sup>14</sup> Gennotte and Leland (1991) demonstrate how market structure can affect information flows in ways that exacerbate shocks to fundamentals.

Mercantile Exchange, transactions involving locals averaged 46 percent of all contracts during the three days prior to the Crash. On October 19, this percentage declined to 31.4 of contract volume and on the following day fell yet again to 24.1 percent.<sup>15</sup> As CME market-making capacity declined, selling shifted from the CME to the New York Stock Exchange overwhelming the NYSE price reporting system. The lack of timely price information added to uncertainty and heightened selling pressure.

An extreme case yes, but the 1987 Crash illustrates linkage between immediacy and creditworthiness. On Black Monday, locals bought 48,487 contracts, selling all but 1,743 before market close. As the price trend over the day was sharply down, most locals ended the day with losses. These losses played an important role in decisions that substantially reduced participation of locals later the same day and the day following. This decline in immediacy at the futures exchange moved selling pressure to New York. Despite this connection, credit extended for market making activities appears to have gone to NYSE specialists. This is not a criticism of the private credit decisions made during this period, but there may be grounds for criticizing an apparent lack of concern for the public's interest. It is fair to question whether credit extended to locals at the futures exchanges might have more effectively served the public interest than the same amount of credit provided to NYSE specialists. Arguably concerns over

---

<sup>15</sup> All of these are substantially less than reported in Curran (2002) in more recent periods (see footnote 6) . These and the following market statistics for October 1987 are drawn from the Report of the Presidential Task Force on Market Mechanisms (Brady report) reprinted in *Black Monday and the Future of*



the viability of the futures clearinghouses precluded this credit allocation. If so, solutions to these problems opens another avenue of response in some future crash. That is to say, a solution to clearinghouse viability problems that puts both stock and futures markets on equal credit standing re-opens the question as to where a credit allocation can best serve the public interest.

A second policy issue is the effect that electronic trading will have. On its face, the cost advantages of electronic trading are so great it is difficult to imagine open outcry having much of a future. However, the immediacy problems noted by Miller (1996) remain unsolved. This researcher believes that open outcry will persist as long as electronic trading operates as a messaging system. Message systems speed up order routing but improving order routing is not sufficient for improving immediacy, orders must also be executed.

Electronic trading must also replicate the immediacy provided by locals in open-outcry markets. When operated as a messaging system every trader is a potential immediacy provider, or not. As is the case with a specialist, immediacy providers face the prospect of trading at an information disadvantage. To succeed, the messaging system must convey information that can be utilized to mitigate this disadvantage. In open outcry markets, successful locals have learned to extract information from market activity. This information is not conveyed by screen-based trading systems. This constrains effective mitigation of the information disadvantage. Absent routes to avoid being disadvantaged,

prospective immediacy providers will require compensation proportional to the level of risk. At times, this cost may deter trading.

This is not to argue that electronic trading can't succeed, these are problems that can be overcome. Open outcry markets developed solutions for immediacy problems prior to their becoming subject to regulatory oversight. Electronic exchanges must achieve similar innovations while satisfying their regulators. At least two problems arise from this need. First, regulators stand to be criticized when negative outcomes are realized, but are not rewarded when outcomes are positive. This payout structure explains the near universality of risk aversion amongst regulators. Second, innovations that threaten the viability of existing competitors motivate attempts to influence regulators for purposes of protecting the status quo. Existing competitors can be expected to provide detailed explanations as to how adoption of proposed innovations may cause regulators to realize the negative outcomes they fear.

Forseeing these difficulties, in 1993 the Chicago Board of Trade proposed a "Pro Markets" approach that segmented market participants by a combination of financial sophistication and capacity to suffer losses. This initiative, itself unsuccessful, ultimately led to adoption of this regulatory philosophy in the Commodity Futures Modernization Act in December 2000. That Act enables the Commodity Futures Trading Commission to effect a multi-tiered regulatory structure. The Commission envisions exchanges operating as either designated contract markets, as derivatives transaction execution facilities (DTFs), or as multilateral transaction execution facilities (MTEFs). The general public is

eligible to participate in designated contract markets, consequently the regulatory oversight of these exchanges is most restrictive. Eligibility at DTFs is limited to cash market participants, that is, traders making or taking delivery. MTEF participation is limited to traders representing institutional firms. Oversight of the latter is minimal, limited to fraud and manipulation concerns. At this point, this regulatory framework appears to remove regulatory impediments that might otherwise deter innovators from solving the immediacy problems noted by Professor Miller. The future of futures appears to one of interesting times.

## References

- Baer, Herbert, Virginia Grace France, James T. Moser, "Opportunity Cost and Prudentiality: An Analysis of Collateral Decisions in Bilateral and Multilateral Settings," Federal Reserve Bank of Chicago Working Paper October 2001.
- Bollen, Nicolas P. B., Tom Smith and Robert E. Whaley, "Optimal Contract Design: For Whom?" working paper, January 25, 2001.
- Curran, John F., "Contract size and Underlying Liquidity," Chicago Mercantile Exchange working paper, February 2002.
- Davidson, Carl, "Equilibrium in Servicing Industries: An Economic Application of Queuing Theory," *Journal of Business* 61(3), 1988, pp. 347-367.
- De Vany, Arthur, "Uncertainty, Waiting Time and Capacity Utilization — A Stochastic Theory of Product Quality," *Journal of Political Economy*, Vol. 84(3), 1976, pp. 523—541.
- De Vany, Arthur and T. R. Saving, "Product Quality, Uncertainty, and Regulation: The Trucking Industry," *American Economic Review*, 1977, pp. 583-594.
- Frech, H. E. III and William C. Lee, "The Welfare Cost of Rationing-By-Queuing Across Markets: Theory and Estimates From the U.S. Gasoline Crises," *Quarterly Journal of Economics* 1987, pp. 97-108.
- Genotte, Gerard and Hayne E. Leland, Market Liquidity, Hedging and Crashes, *American Economic Review* December 1990, pp. 999-1021.
- Kyle, Albert S. and Terry A. Marsh, "On the Economics of Securities Clearing and Settlement," Working Paper 1994.
- Miller, Merton H., "The Future of Futures," *Derivatives and Public Policy: Proceedings of a Conference*, Federal Reserve Bank of Chicago, June 1996.
- Miller, Merton H. and Daniel Orr, "A Model of the Demand for Money by Firms," *Quarterly Journal of Economics* 80, 1966, pp. 413-435.
- Miller, Merton H. and Daniel Orr, "The Demand for Money by Firms: Extensions of Analytic Results," *Journal of Finance* 23(5), 1968, pp. 735-759.
- Moser, James T., "Pricing Futures Contracts: Restrictions on Trading-Day Price Changes", Doctoral Thesis, Ohio State University, 1986.
- Moser, James T. "Origins of the Modern Exchange Clearinghouse: A History of Early Clearing and Settlement Methods at Futures Exchanges," *Classic Futures*:

*Lessons from the Past for the Electronic Age*, edited by Lester G. Telser. Risk Publications: Chicago 2000, pp. 277-319.

Moser, James T. "Immediacy, Credit Risk and Exchange Organization," Federal Reserve Bank of Chicago *Working Paper March 2002*.

Naor, P., "The Regulation of Queue Size by Levying Tolls," *Econometrica* 37(1), 1969, pp. 15-23.

Silber, William L. "Marketmaker Behavior in an Auction Market: An Analysis of Scalpers in Futures Markets," *Journal of Finance* 39, no. 4, September 1984 pp. 937-953.

Telser, Lester G. "Futures and Actual Markets: How They Are Related," *Journal of Business* 59(2), part 2, 1986, pp. S5-S20.