ELSEVIER

Journal of Financial Markets 4 (2001) 33–71

Journal of FINANCIAL MARKETS

www.elsevier.nl/locate/econbase

# A simple model of payment for order flow, internalization, and total trading cost☆

Robert Battalio[a], Craig W. Holden[b,*]

[a]*Mendoza College of Business Administration, University of Notre Dame, Notre Dame, Indiana 46556, USA*

[b]*Kelley School of Business, Indiana University, Bloomington, IN 47405, USA*

**Abstract**

We show that *externally-verifiable characteristics* (inexpensive for a third-party to verify) of traders or orders allow profitable purchasing of order flow and internalization. We introduce *total trading cost*, defined as the effective half spread plus the broker's per share commission, as a measure of execution quality. We use this measure to reinterpret prior empirical studies of: (1) execution quality across trading venues and (2) cream-skimming by purchasers of order flow. Finally, we show brokers can use their direct relationships with customers to assess *internally-verifiable characteristics* (inexpensive for direct verification) in order to increase profits extracted from customer orders. 

*JEL classification:* G14; G24

*Keywords:* Microstructure; Payment for order flow; Internalization; Total trading costs; Decimalization

☆We thank Jeff Bacidore, Mitch Berlin, Hendrik Bessembinder, Tarun Chordia, Paul Clyde, Mark Flannery, Robert Jennings, Maureen O'Hara, Eric Rasmusen, Richard Rosen, Robert Schwartz, and seminar participants at the Western Finance Association, Indiana University Biennial Symposium, the Rutgers University Conference on Financial Economics and Accounting, the Ohio State Conference on Dealer Markets, and Georgia State University for helpful suggestions and/or discussions. We thank Len Lundstrum and Andy Waisburd for quality research assistance. We are solely responsible for any errors.

* Corresponding author. Tel.: 812-855-3383; fax: 812-855-5875.

*E-mail address:* cholden@indiana.edu (C.W. Holden).

1386-4181/01/$ - see front matter 
PII: S1386-4181(00)00015-X

## 1. Introduction

Two major developments in the competition for securities trading are: (1) *payment for order flow*, the payment of cash inducements to brokers to obtain customer orders, and (2) *internalization*, the execution of a customer order against a broker's own account. Both developments, which involve the execution of customer orders away from the primary market at prices which are no worse than the best prices quoted in the market, have substantially increased the migration of order flow away from primary markets. For example, one major purchaser of order flow, third-market dealer Bernard L. Madoff Investment Securities (Madoff), executed over 10 percent of all orders in New York Stock Exchange (NYSE)-listed securities in 1991.[1] As another example, Battalio, Greene and Jennings (BGJ) (1997) estimate that 7.0 percent of all orders, in a set of actively-traded, NYSE-listed securities, were internalized on the Cincinnati Stock Exchange (CSE) during September 1994.[2] BGJ also find that the Boston Stock Exchange doubled its trade volume after it introduced a program in July 1994 that increased the opportunities for brokers to internalize orders.

Previous literature and popular accounts have suggested that both developments are driven by tick size rules that restrict prices to be on a coarse grid (i.e., $\$\frac{1}{8}$ increments).[3] This implies that both developments will be eliminated when prices are permitted to be on a fine grid, such as the current plan to switch U.S. equity markets to *decimal trading* (i.e., one cent increments) before the end of 2000.[4]

The prediction that payment for order flow and internalization will be eliminated by a finer tick size has been rejected by recent empirical evidence. Direct evidence is offered by both Porter and Weaver (1997) and Ricker (1997), who find no reduction in internalization following the reduction of the minimum trading variation on the Toronto Stock Exchange from \$0.125 to \$0.05.[5] Indirect evidence is provided by Ahn et al. (1996), who find no change in the

[1] Forbes, 6 January, 1992.

[2] Based on the calculation that 8.94% of all orders in a set of 256 NYSE-listed securities eligible for preferencing under the Cincinnati Preferencing Pilot were executed on the Cincinnati Stock Exchange (CSE) and 78% of all orders executed on the CSE were internalized.

[3] Chordia and Subrahmanyam (1995) formalize this intuition by developing a model in which 'the practice of payment for order flow and the possibility of inferior execution can arise naturally in response to a finite tick size'. Cheng (1995) develops a tick size driven model of payment for order flow. Kandel and Marx (1999) develop a tick size driven model of payment for order flow and preferencing in the NASDAQ context.

[4] For additional information, see the January 2000 issue of NYSE's newsletter, 'The Exchange'.

[5] The 'minimum trading variation' is the smallest difference that is permitted between buying and selling prices. See Harris (1997) for an excellent review of this evidence and other evidence about the impact of decimalization.

percentage of trades executed away from the floor of the American Stock Exchange (much of which is based on payment for order flow and internalization) when it reduced its tick size for low-priced stocks from $\$\frac{1}{8}$ to $\$\frac{1}{16}$.[6]

This raises the question, why does finer tick size trading fail to eliminate payment for order flow and internalization? We address this puzzle by developing a simple model of payment for order flow and internalization in a world where prices are permitted to be on an infinitely fine grid (the real line). Previous theoretical models of payment for order flow[7] were based on *anonymous trading*, where dealers cannot observe any characteristic of the trader or order (other than order size) which is helpful in distinguishing which orders are more likely to be informed than others. By contrast, we develop a simple model based on *externally-verifiable characteristics*[8] of the trader and/or order that are helpful in identifying what orders are more likely to be informed.

The first contribution of our paper is to show that externally-verifiable characteristics are sufficient to support the *existence* of: (1) payment for order flow and (2) internalization — even with prices permitted to be on the real line and competitive primary dealers. The intuition for this result is that externally-verifiable characteristics permit a *sorting* of orders based on the likelihood of information content. Purchasers and internalizers use these characteristics when deciding which orders to interact with. For example, Madoff is very insistent on not purchasing any orders from professional traders, but is happy to purchase orders from nonprofessional traders.[9,10] Assuming that professional traders are more likely to be informed than nonprofessionals, purchasers are profiting by *selectively purchasing* nonprofessional orders and having the professional orders

[6] See Ahn et al. (1996, footnote 15).

[7] Glosten (1991), Cheng (1995), Chordia and Subrahmanyam (1995), Easley et al. (1996), and Kandel and Marx (1999).

[8] We use the word 'externally' here to mean external to the relationship between the broker and the trader. Externally-verifiable characteristics are objective characteristics of the traders or orders that are inexpensive for a third-party to verify.

[9] A 1991 report of the NASD Payment for Order Flow Committee confirms that it is standard practice for purchasers of order flow to exclude orders from professional traders and to exclude program trades.

[10] Conversations with purchasers of order flow indicate that both ex-ante and ex-post evaluations are used to decide whether order flow is professional or not. For example, to minimize trading costs, professional traders typically place orders with deep discount brokerage houses. As a result, some purchasers refuse to purchase order flow from deep discount brokerage houses. After purchasers begin receiving order flow from brokerage houses, they periodically monitor for trades that appear to be professional. For example, brokers who route a series of large orders in the same security (i.e., broken up block trades) and/or who systematically route orders that arrive as the depth at a quote is being depleted (i.e., order flow from momentum or day traders) are identified as brokers who are routing professional order flow. If these brokers do not stop routing these orders to the purchaser, the purchaser severs the order purchasing arrangement.

passed through to the primary dealers.[11] Alternatively, a broker could *selectively internalize* nonprofessional orders and route professional orders to the primary dealers. Our model provides a simple, direct explanation of payment for order flow or internalization where prices are not conditioned on all factors correlated with order flow profitability (i.e., prices may be posted as a function of order size but not trader type). In order to reject this 'null hypothesis', other, more complex, models would need to explain as much and more.

Our second contribution is to define a new measure of the cost of trading called Total Trading Cost ($TTC$). We build on the concept of Effective Half Spread ($EHS$),[12] which is a standard measure of the cost of trading in the literature. $EHS$ is defined as the transaction price minus the quote midpoint for market buy orders and as the quote midpoint minus the transaction price for market sell orders. The quote midpoint is the simple average of the bid and ask. Our new measure, the total trading cost ($TTC$) is defined as $EHS +$ (the broker's dollar amount per share commission). $TTC$ is the entire cost of trading faced by an outside trader. It is very important to include the broker's dollar amount per share commission when calculating this entire cost of trading, because there are differences across equilibria in the size of the payment for order flow and in what gets passed on to the uninformed trader. Hence, $TTC$ is the only way to compare the *whole* cost of trading across equilibria.[13]

To illustrate the usefulness of our new measure $TTC$, we analyze the evolution of payment for order flow over time and are able to resolve a conflict in the empirical literature. Madoff innovated payment for order flow in the early eighties by offering brokers one to two cents per share for market orders in certain NYSE-listed securities. Madoff's volume of activity has grown steadily since then. Late in the 1980s Madoff's monopoly on purchasing order flow was challenged by a variety of 'Madoff knock-offs'. For example, D.E. Shaw paid two cents per share for market orders in those securities which comprise the S&P 100 in the mid-1990s. Battalio et al. (2000) report that Trimark Securities paid an average of 1.3 cents per share for market orders in S&P 100 issues. In

[11] Glosten (1991) shows that monopoly power by the primary dealer is an alternative explanation of payment for order flow. By contrast, we are able to show that purchase of order flow is possible even with competitive primary dealers and prices permitted to be on the real line.

[12] See Huang and Stoll (1996). Also, called the 'liquidity premium' in some papers.

[13] We also examine primary dealers' ex-ante profits, which are equal to the expected value of the Realized Half Spread. (E[$RHS$]). Following the convention of Huang and Stoll (1996), the (E[$RHS$]) is defined as the expected value of the terminal value minus the transaction price for market buy orders or as the transaction price minus the expected value of the terminal value for market sell orders. In the absence of finite ticks and in the presence of either competitive dealers or of alternative liquidity suppliers (purchasers or internalizers), primary dealers generally earn zero (E[$RHS$]). The only exception that we consider is a monopolistic primary dealer with no purchasing or internalization, who earns positive (E[$RHS$]).

summary, the financial markets have evolved over time through three stages: (1) nonexistent purchasing, (2) monopolistic purchasing by the original entrepreneur, and (3) relatively competitive purchasing by multiple entrants.[14]

We analyze these three stages by tracing patterns over time in two dimensions: (1) the $TTC$ and (2) the difference in the probability of informed trading *on* the primary market vs. *away* from the primary market. We use this analysis to shed light on the controversial issue of whether payment for order flow and internalization are 'cream-skimming' or cost competition. Recent empirical studies claim opposite results on this issue.

On the first dimension, Battalio (1997) examines the *EHS* before and after Madoff begins purchasing order flow in a given stock. He finds that the *EHS* does not change when Madoff enters the market.[15] If we assume that the broker does *not* pass any of the payment for order flow through to the customer, then the $TTC$ stays constant. Alternatively, if we assume that the broker *does* lower commissions, then the $TTC$ decreases. The large-scale emergence of deep-discount commissions by on-line brokers that accept payment for order flow (e.g., E-trade) suggests that some of the payment is being passed through to the customer to attract orders,[16] and thus, the $TTC$ is lower. Our new measure $TTC$ picks up the drop in the cost of trading, whereas the *EHS* does not. A lower $TTC$, as shown by Battalio's evidence combined with deep-discount commissions by on-line brokers that accept payment for order flow, suggests that payment for order flow is *not* cream-skimming.

On the second dimension, Easley et al. (EKO) (1996) use data on buy/sell imbalances to estimate the probability of informed trading on the NYSE vs. the CSE (where internalized and purchased orders are frequently executed). EKO find that 'there is a significant difference in the information content of orders executed in New York and Cincinnati'. They conclude 'that this difference is consistent with cream-skimming'.[17]

We resolve this seeming conflict by showing that our theoretical model can generate *both* empirically-observed results simultaneously. Specifically, our model can generate: (1) a *drop* in the $TTC$ and (2) separating equilibria with

[14] Similarly, internalization has grown steadily over the years. It has grown to the point where BGJ find that 71 percent of small size trades in a set of NYSE-listed securities were executed, most likely by purchasers and internalizers of order flow, on trading venues other than the NYSE.

[15] BGJ, Neal and Reiffen (1994), and Lamoreux and Schnitzlein (1997) also find evidence that the diversion of order flow away from the primary market does not increase the *EHS*.

[16] As of June 2000, Ameritrade unit Freetrade.com began offering free internet trading to market order traders. A June 19, 2000 Wall Street Journal article noted that payment for order flow is an important revenue that makes it possible to offer free trades.

[17] Bessembinder and Kaufman (1997), Chordia and Subrahmanyam (1995), and Lin et al. (1995) also find empirical evidence which suggests that a significant portion of the order flow in NYSE-listed securities which is diverted away from the NYSE is informationless.

a *lower* probability of informed trading away from the primary market. We conclude that the evidence of Battalio (1997) and of EKO are both consistent with our segmentation theory and are therefore consistent with each other.

The third contribution of our paper is to develop a further explanation of internalization. We distinguish between externally-verifiable characteristics and internally-verifiable characteristics. Externally-verifiable characteristics are objective characteristics that can be verified inexpensively by a party outside of the broker/trader relationship (e.g., order size, professional vs. nonprofessional, program vs. non-program, etc.). Internally-verifiable characteristics are characteristics derived from the broker/dealer relationship which are either subjective and/or prohibitively expensive for a third-party to verify (i.e., sophisticated traders vs. naïve traders, active traders vs. occasional traders, trader wealth, subjective indicators of trader motivation, etc.). This distinction is crucial because third-market dealers do not have a direct relationship with traders, but must instead contract with brokers to obtain order flow. Hence, third-market dealers can selectively purchase order flow based on externally-verifiable characteristics, but not internally-verifiable characteristics. By contrast, brokers have direct relationships and personal contacts with traders. This allows brokers to use internally-verifiable characteristics of traders and/or orders to perform an additional sorting of the order flow. We show that brokers can profit by internalizing the additional sorts that are less likely to be informed, even with prices on the real line. Hence, internalizers have an advantage over purchasers.

The fourth contribution of our paper is to revisit the empirical literature on trading costs on and away from the NYSE. We examine 13 comparisons from five different papers, all of which use Effective Half Spread (*EHS*) as the basis of comparison. 11 of these 13 *EHS* comparisons favor the NYSE and only 2 favor the non-NYSE venue. We define the $TTC$-Equalizing Passthrough Percentage as the percentage of the payment for order flow that the broker must passthrough to the trader in order to equalize the $TTC$ of both venues. Of the 11 comparisons that initially favored the NYSE, two of them yield $TTC$-Equalizing Passthrough Percentages greater than 100 percent and thus cannot be reversed. The remaining 9 *might* be reversed depending on the degree of competition of the brokerage industry. For example, if the brokerage market is sufficiently competitive that brokerage firms would be forced to passthrough over 50 percent of the payment for order flow, then 5 of the remaining 11 comparisons would be reversed. That is, the majority of the comparisons (7 out of 13) would now favor the non-NYSE venue over the NYSE. We suggest that future papers on performance by venue calculate the $TTC$-Equalizing Passthrough Percentage (which does not require brokerage data) in order to investigate the robustness their conclusions to the impact of payment for order flow and internalization on brokerage commissions.

The plan of the paper is as follows. Section 2 develops a model of the purchase of order flow using externally-verifiable characteristics and shows that

purchasing can take place even when prices are permitted on the real line. Then it develops an internalization model and shows that internalization can take place using externally-verifiable characteristics with prices on the real line. Section 3 uses the concept of total trading cost to analyze the evolution of payment for order flow over time and resolve an empirical conflict. In addition, it provides a numerical calibration that illustrates different equilibria. Section 4 expands the internalization model to internally-verifiable characteristics. Section 5 revisits the empirical literature on the performance of different trading venues by calculating the $TTC$-equalizing passthrough percentage. Section 6 concludes. All proofs and an analytic solution in a special case are in the appendix.

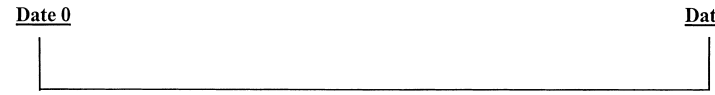
## 2. Both activities using externally-verifiable characteristics

### 2.1. A payment for order flow model

We extend the Glosten and Milgrom (1985) model to incorporate payment for order flow and order-routing by brokers. There exists a single risky asset, which is listed on the primary exchange, and a riskfree asset. The riskfree rate is normalized to zero. The risky asset has a terminal value of $v$ from the interval $[v_L, v_H]$ with an unconditional mean $E[v]$. There are two sides of the market: the (trader) buy side and the (trader) sell side. Let $q$ be the quote midpoint, which is defined as the simple average of the buy side quoted price (the ask) and the sell side quoted price (the bid). We wish to connect to the Huang and Stoll (1996) definition of the $EHS$, which uses $q$ as the pre-trade estimate of the risky asset's true value. In order to formally justify using $q$ to calculate $EHS$, we assume that the two sides of the market are symmetric.[18] This immediately generates the result that $q = E[v]$ and thus provides the basis for using $q$ to calculate $EHS$ and $TTC$.

The two sides of the market are disconnected from each other. Without loss of generality, we focus the exposition on the buy side. The sell side is incorporated by analogy. There are five classes of risk neutral, economic agents: primary dealers, third-market dealers ('purchasers'), brokers, professional traders, and nonprofessional traders. Both professional and nonprofessional traders may be either informed or uninformed. We adopt the convention that professional traders are more likely to be informed than nonprofessional traders. The sequence of actions these agents take on the trader buy — dealer sell side is explained as a timeline in Fig. 1.

[18] Specifically, we assume that: (1) $v$ is symmetrically distributed about $E[v]$, (2) uninformed traders have equal probabilities of submitting buy and sell orders, and (3) the arrival probabilities of each type of trader are the same on both sides. Implicitly, this is what Huang and Stoll (1996) are assuming.




Step 1: The $i$th primary dealer quotes an ask $a^i$ for both pro and non-pro buys.

Step 2: The $j$th purchaser offers a payment $p^j$ for the contractual set $c^j$ of buys.

Step 3: The $b$th broker sets a negative dollar amount per share commission $\left(c^b < 0\right)$ -- that is, sets a dollar amount per share rebate -- for both pro and non-pro buys.

Step 4: A trader arrives, chooses to submit an unit market buy or do nothing, chooses a broker to handle the order, sends the order to a broker, and receives the dollar amount per share rebate. With probability $\beta$, the trader type is professional ($t = pro$). Then:

- With probability $\alpha^{pro}$, the trader is informed and chooses whether to buy $\left(x^{pro} = 1\right)$ or not $\left(x^{pro} = 0\right)$.
- With probability $1-\alpha^{pro}$, the trader is uninformed and a random number of shares $u^{pro} \in \{0,1\}$ is drawn.

Alternatively with probability $1-\beta$, the trader type is non-professional ($t = non$). Then:

- With probability $\alpha^{non} < \alpha^{pro}$, the trader is informed and chooses whether to buy $\left(x^{non} = 1\right)$ or not $\left(x^{non} = 0\right)$.
- With probability $1-\alpha^{non}$, the trader is uninformed and a random number of shares $u^{non} \in \{0,1\}$ is drawn.

Step 5: A broker who receives a market buy:

- observes the trader type (pro vs. non-pro)
- chooses whether to send it to the primary dealer or to the purchaser.

Later in the paper (in subsection 2.3), the broker may choose to internalize it.

Step 6: The market buy is executed.

The security's terminal value $v$ is realized.

Fig. 1. The model timeline.

Referring to the figure, on date 0 step 1 each primary dealer quotes an ask price that would apply to any market buy order whether it comes from a professional or nonprofessional trader. On date 0 step 2, each purchaser offers a dollar payment amount to brokers in return for brokers routing a contractual set of orders to them. We assume that purchasers can externally-verify whether orders come from professional traders or nonprofessional traders. So purchasers might specify in the contractual set that they are willing to pay for: (1) non-professional orders only, (2) professional orders only, (3) both types of orders, or

(4) neither. On date 0 step 3, each broker sets a negative[19] dollar amount per share commission – that is, sets a dollar amount per share rebate — for any market buy order, whether it comes from a professional or nonprofessional trader. Each broker has a fiduciary responsibility to the trader to verify that whomever executes a market buy will do so at a price which is not above the lowest ask price available in the market. On date 0 step 4, a trader arrives, chooses whether to submit an unit market buy order or do nothing, chooses a broker to handle the order, sends the order to a broker and receives the dollar amount per share rebate. On date 0 step 5, a broker who receives a market buy observes the trader type and chooses whether to send it to the primary dealer or to the purchaser. The broker can only send it to the purchaser if the trader type (pro vs. nonpro) is in the contractual set that the purchaser is willing to accept. On date 0 step 6, the order is executed by whichever dealer it was routed to in step 5. On date 1, the security's terminal value is realized.

Table 1 provides an overview of the choice variables and random variables in the model. Panel A summarizes the choice sets of the endogenous agents. The first column lists the four classes of endogenous agents. For each class of agent, the choice variables and the space of the choice variables are listed. Panel B summarizes the order submission random variable of the exogenous agent, namely the uninformed trader.

For convenience, we place additional minor restrictions on the distribution of $u$ and $v$. Specifically, we assume:

- $u$ and $v$ are independent of each other,
- $\Pr(u^t = 1) > 0$ for $t = pro, non$, and
- $v$ has positive mass at two or more points.

### 2.2. Solving the model

We start at the end and work backwards. Thus, we start at step 5 (see Fig. 1), where the broker decides whether to send the received market buy order to a primary dealer or to a purchaser. At this point in the timeline, the first four steps have already happened and are thus sunk decisions. In particular, the broker paid the rebate to the trader in step 4, so the only remaining decision is where to route the order. If any of the purchasers is offering a positive payment and the primary dealers are offering nothing, it is clearly optimal to send the order to a purchaser. If all of the purchasers and all of the primary dealers are

[19] The commission is negative simply because we do not model a variable cost of providing brokerage services. If we did model a variable cost of sufficient size, then the commission would be positive. In the real world, the 'rebate' that we have in mind are lower commissions than would be charged in the absence of payment for order flow or internalization.

Table 1
Overview of choice variables and random variables in the model[a]

| Class of agents | Choice variables | Choice variable space |
|---|---|---|
| *Panel A: The choice sets of the endogenous agents* | | |
| Primary dealers | $a^i =$ ask of the $i$th primary dealer | $[v_L, v_H]$ |
| Third-market dealers | $p^j =$ payment of the $j$th third-market dealer | $[0, \infty)$ |
| | $s^j =$ contractual set of the $j$th third-market dealer | $\{none, pro, non, both\}$ |
| Brokers | $c^b =$ per share commission (rebate) charged by the $b$th broker | $(-\infty, 0]$ |
| Informed trader | $x^t =$ buy shares from the informed trader of type $t$ | $\{0, 1\}$ |
| *Panel B: Order submission by the exogenous agent* | | |
| Agent | Random variable | Distribution support |
| Uninformed trader | $u^t =$ buy shares from the uninformed trader of type $t$ | $\{0, 1\}$ |

[a] $v_L$ is the lower bound of $v$ and $v_H$ is the upper bound of $v$.

offering nothing, then the broker is indifferent as to where the order goes. When the broker is indifferent, we adopt the tie-breaking convention that orders are sent to a primary dealer.

Next, work backwards to Step 4, the informed trader's problem. The informed trader observes the terminal value of the risky asset $v$, where $v$ is a real value from the bounded interval $[v_L, v_H]$ with an unconditional mean $E[v]$. Focusing on the market buy side only, the informed trader can buy the risky asset for a total price equal to the quote midpoint $q$ plus the lowest of our new measure the total trading cost ($TTC$). The $TTC$ is the sum of the ask price $a$ plus the dollar amount per share commission $c$. The informed trader of type $t$ ($t \in \{pro, non\}$) chooses whether to buy ($x^t = 1$) or not ($x^t = 0$) in order to maximize expected profits

$$\underset{x^t}{\text{Max}}\ E[x^t\{v - (q + TTC)\} \mid v] = \underset{x^t}{\text{Max}}\ x^t\{v - (q + TTC)\}. \tag{1}$$

As is well known, the optimal strategy is bang–bang

$$x^t = \begin{cases} 1 & \text{when } v \geqslant q + TTC, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Hence, the informed market buy ($x^t$) will be submitted to a broker charging the lowest commission $c$ (e.g., offering to pay the largest rebate). We assume that the

uninformed market buy ($u^t$) will also be submitted to a broker offering the lowest commission $c$.

Next we roll back to Step 3 of the timeline, which is the broker's problem. Each broker knows that when the next buy order arrives, it will be possible to determine the type $t$ of the order submitter. The $b$th broker charges a negative per share commission $c^b$ on all buy orders and will receive payment $p^j$ from the $j$th purchaser for all orders in the contracted set $s^j$ and no payment otherwise. The $b$th broker will choose $c^b$ to maximize expected profits

$$\underset{c^b}{\text{Max}}\ (f^j)\text{E}[(p^j + c^b) \mid t \in s^j] + (1 - f^j)\text{E}[(c^b) \mid t \notin s^j], \tag{3}$$

where $f^j$ is the probability that the next buy order is of type $t \in s^j$ and $1 - f^j$ is the probability that the next buy order is of type $t \notin s^j$. The variables $p^j$ and $c^b$ are not random and can be pulled out of the expectations operator. In equilibrium, there will be a single, competitive per share commission $c$ that earns zero expected profits

$$f^j(p^j + c) + (1 - f^j)c = 0. \tag{4}$$

Solving for $c$, we obtain

$$c = -f^j p^j. \tag{5}$$

In this subsection, we analyze *competitive* primary dealers and *monopolistic* purchasing and use the double subscript cm to denote this. Rolling back to Step 2 of the timeline, we analyze the monopolistic purchasing problem. The monopolistic purchaser knows that primary dealers will execute any order at the competitive ask $a_{\text{cm}}$. After paying $p_{\text{cm}}$, the monopolistic purchaser ends up with a net price $a_{\text{cm}} - p_{\text{cm}}$ in order to sell shares. The monopolistic purchaser chooses $p_{\text{cm}}$ and $s_{\text{cm}}$ to maximize expected profits

$$\underset{p_{\text{cm}}, s_{\text{cm}}}{\text{Max}}\ (f_{\text{cm}})\text{E}[(a_{\text{cm}} - p_{\text{cm}}) - v \mid t \in s_{\text{cm}}], \tag{6}$$

where $f_{\text{cm}}$ is the probability that the next buy order is of type $t \in s_{\text{cm}}$. Since the monopolist's expected profits are strictly decreasing in the amount of payment $p_{\text{cm}}$, the optimal policy is to set the payment equal to an arbitrarily small amount $\varepsilon > 0$.

Finally we roll back to Step 1 of the timeline, which is the competitive primary dealer's problem. The only potential orders that primary dealers receive are orders that are not in the contractual set. The $i$th primary dealer chooses an ask $a^i$ to maximize expected profits

$$\underset{a^i}{\text{Max}}\ \text{E}[a^i - v \mid t \notin s_{\text{cm}}, TTC_{\text{cm}}]. \tag{7}$$

In equilibrium, the competitive ask $a_{\mathrm{cm}}$ earns zero expected profits

$$a_{\mathrm{cm}} - \mathrm{E}[v \mid t \notin s_{\mathrm{cm}}, TTC_{\mathrm{cm}}] = 0. \tag{8}$$

Eq. (8) can be solved for the competitive ask $a_{\mathrm{cm}}$, which allows the primary dealer to break even on professional orders only. At this price, the monopolistic purchaser would loose money if he tried to purchase professional orders. Hence, the optimal contracted set is to purchase nonprofessional orders only ($s_{\mathrm{cm}} = non$) for an arbitrarily small amount $\varepsilon > 0$. The resulting equilibrium is described in the following proposition.

*Proposition 1. A payment for order flow model based on*: (i) *prices on the real number line*, (ii) *competitive primary dealers*, *and* (iii) *monopolistic purchasing* (cm) *yields a separating monopolistic* (sm) *equilibrium as characterized by*

$$p_{\mathrm{cm}} = \varepsilon, \tag{9}$$

$$s_{\mathrm{cm}} = non, \tag{10}$$

$$c_{\mathrm{cm}} = -\Pr(t \in non \mid buy, TTC_{\mathrm{cm}})\varepsilon, \tag{11}$$

$$\begin{aligned} a_{\mathrm{cm}} &= \mathrm{E}[v \mid t \in pro, TTC_{\mathrm{cm}}] \\ &= w_{\mathrm{cm}}\mathrm{E}[v \mid v \geqslant q + TTC_{\mathrm{cm}}] + (1 - w_{\mathrm{cm}})\mathrm{E}[v] \equiv a_{\mathrm{sm}}, \end{aligned} \tag{12}$$

$$x_{\mathrm{cm}} = \begin{cases} 1 & when\ v \geqslant q + TTC_{\mathrm{cm}}, \\ 0 & otherwise, \end{cases} \tag{13}$$

$$q = \mathrm{E}[v], \tag{14}$$

$$EHS_{\mathrm{cm}} = a_{\mathrm{cm}} - q, \tag{15}$$

$$TTC_{\mathrm{cm}} = EHS_{\mathrm{cm}} + c_{\mathrm{cm}} \equiv TTC_{\mathrm{sm}}, \tag{16}$$

$$\mathrm{E}[RHS_{\mathrm{cm}}] = 0, \tag{17}$$

*where* $w_{\mathrm{cm}}$ *and* $\Pr(t \in non \mid buy, TTC_{\mathrm{cm}})$ *are given in the appendix.*

Proposition 1 demonstrates an equilibrium in which payment for order flow takes place even though prices are permitted to be on the real line and primary dealers are competitive. Indeed, the monopolistic purchaser is earning positive rents! The monopolistic purchaser's profit margin is the entire difference in expectation between the two types of orders ($\mathrm{E}[v \mid t \in pro, TTC_{\mathrm{cm}}] - \mathrm{E}[v \mid t \in non, TTC_{\mathrm{cm}}]$) less an arbitrarily small amount $\varepsilon$. This result is driven by the fact that there is greater adverse selection in professional buys than in nonprofessional buys. Thus, if the separating monopolistic ask $a_{\mathrm{sm}}$ is set to break even on professional buys, then there is plenty of profit opportunity in

nonprofessional buys. The monopolistic purchaser simply has to offer a token payment to purchase these nonprofessional buys. This result stands in sharp contrast to the absence of payment for order flow in a tick-driven model with competitive dealers when the tick size is set equal to zero.

It is interesting to note that the *TTC* for trades executed by purchaser in the third market is less than the *TTC* in the primary market (by an arbitrarily small amount), even thought the *EHS* in the third market is equal to the *EHS* in the primary market. In other words, the *TTC* metric can change the performance ranking that comes from the *EHS* metric. We will find this is true in all of our separating equilibria. The primary dealers' expected profit, which is the expected value of the Realized Half Spread E[*RHS*] (see footnote 13), is zero.

### 2.3. An internalization model

In this subsection we create a model of internalization, which is very similar to the purchase of order flow model based on externally-verifiable characteristics. Later in Section 4, we will extend this model to allow brokers to use their relationships and personal contacts with customers to sort traders and their order flows using internally-verifiable characteristics (characteristics derived from the broker/dealer relationship which are either subjective and/or prohibitively expensive for a third-party to verify). We show that this is profitable and thus there is a key distinction between internalization and the purchase of order flow. Brokers internalize orders using *both* externally-verifiable and internally-verifiable characteristics. Whereas, purchasers are limited using *externally-verifiable characteristics only*, since they do not have any direct contact with the investor.

In this subsection, each dealer has the choice to send the order to the primary dealers or to internalize it (i.e., execute it on the broker's own account). For simplicity, we drop the step involving the purchase of order flow (Step 2). Fig. 2 updates the timeline for the modified model.

We start at the modified Step 5 of the timeline, where the broker's decision is whether to send the received market order to the primary dealer or internalize it. Very simply, the broker's optimal policy is to internalize an order when the expected terminal value conditional on that type of order is less than the ask $a$ and pass it through to the primary dealer otherwise.

For the modified Step 3, we add notation for the broker's internalization decision (see Table 2).

In Step 3, the $b$th broker anticipates making a profit margin of $\pi^b$ on all orders in the internalization set $i^b$ and no profit margin otherwise. The $b$th broker will choose $c^b$ and $i^b$ to maximize expected profits

$$\underset{c^b, i^b}{\text{Max}} \; (f^b)\mathrm{E}[(\pi^b + c^b) \mid t \in i^b] + (1 - f^b)\mathrm{E}[(c^b) \mid t \notin i^b], \tag{18}$$

**Date 0**

Drop Step 2

Modified Step 3: The $b$th broker offers a commission $c^b$ for all buy orders.

Modified Step 5: The broker receiving the market buy:
- observes the trader type (professional vs. nonprofessional)
- updates his conditional expectation of the terminal value, and
- chooses whether to send it to the primary dealer or internalize it.

Fig. 2. Changes in the model timeline for internalization based on externally-verifiable characteristics.

Table 2
Additional choice variable for internalization based on externally-verifiable characteristics[a]

| Class of agents | Choice variables | Space of the choice variable |
|---|---|---|
| Brokers | $i^b$ = internalization set of the bth broker | $\{none, pro, non, both\}$ |

[a]For example, $i^b = non$ means internalize nonprofessional orders.

where $f^b$ is the probability that the next buy order is of type $t \in i^b$ and $1 - f^b$ is the probability that the next buy order is of type $t \notin i^b$. In equilibrium, there will be a single, competitive commission $c$, probability $f$ of receiving a buy order in the competitive internalization set $i$, and competitive profit margin $\pi$. As before, we set the broker's expected profits to zero and solve for $c$ to obtain $c = -f\pi$.

Step 1 proceeds as in Section 2.2. Let $m$ be the conditional mean of the terminal value for nonprofessional orders. The internalization model yields the following equilibrium.

*Proposition 2. When there are competitive primary dealers and internalization by competitive brokers* (*ci*), *the resulting equilibrium is separating competitive* (*sc*) *as characterized by*

$$i_{\mathrm{ci}} = non, \tag{19}$$

$$c_{\mathrm{ci}} = -\Pr(t \in non \mid buy, TTC_{\mathrm{ci}})(a_{\mathrm{sc}} - m_{\mathrm{sc}}), \tag{20}$$

$$\begin{aligned} a_{\mathrm{ci}} &= \mathrm{E}[v \mid t \in pro, TTC_{\mathrm{ci}}] \\ &= w_{\mathrm{ci}}^{pro}\mathrm{E}[v \mid v \geqslant q + TTC_{\mathrm{ci}}] + (1 - w_{\mathrm{ci}}^{pro})\mathrm{E}[v], \end{aligned} \tag{21}$$

$$\begin{aligned} m_{\mathrm{ci}} &= \mathrm{E}[v \mid t \in non, TTC_{\mathrm{ci}}] \\ &= w_{\mathrm{ci}}^{non}\mathrm{E}[v \mid v \geqslant q + TTC_{\mathrm{ci}}] + (1 - w_{\mathrm{ci}}^{non})\mathrm{E}[v], \end{aligned} \tag{22}$$

$$x_{\mathrm{ci}} = \begin{cases} 1 & \textit{when } v \geqslant q + TTC_{\mathrm{ci}}, \\ 0 & \textit{otherwise}, \end{cases} \tag{23}$$

$$q = \mathrm{E}[v], \tag{24}$$

$$EHS_{\mathrm{ci}} = a_{\mathrm{ci}} - q, \tag{25}$$

$$TTC_{\mathrm{ci}} = EHS_{\mathrm{ci}} + c_{\mathrm{ci}} \equiv TTC_{\mathrm{sc}}, \tag{26}$$

$$\mathrm{E}[RHS_{\mathrm{ci}}] = 0, \tag{27}$$

*where* $w_{\mathrm{ci}}^{pro}$, $w_{\mathrm{ci}}^{non}$, *and* $\Pr(t \in non \mid buy, TTC_{\mathrm{ci}})$ *are given in the appendix.*

Proposition 2 demonstrates an equilibrium in which internalization takes place even though prices are permitted to be on the real line and primary dealers are competitive. Primary dealers set the ask equal to the conditional mean given professional orders, since they end up with professional orders only. Brokerage commissions consist of a payment equal to the probability of receiving a non-professional order $\Pr(t \in non \mid buy, TTC_{\mathrm{sc}})$ times the difference in conditional means $(a_{\mathrm{sc}} - m_{\mathrm{sc}})$. Essentially, competitive internalizers compete away all of their rents. A separating competitive equilibrium also obtains when there are competitive purchasers of order flow and no internalization (see Proposition 3). In either case, the rents generated by selectively executing orders are returned to traders in the form of lower brokerage commissions (i.e., higher rebates). The *TTC* in the third market is less than the *TTC* in the primary market (by a finite amount), even though the *EHS* in the third market is equal to the *EHS* in the primary market.[20] As before, the *TTC* metric changes the performance ranking that comes from the *EHS* metric. In Proposition 4, we will show that this dissipation of rents implies that the *TTC* in a market with competitive internalizers or competitive purchasers equals the *TTC* in a market where all orders are executed by competitive dealers.

This section has shown that it is possible to get either purchase of order flow or internalization in a market with competitive primary dealers and no finite ticks. The contrast between the two equilibria in terms of the capture of rents demonstrates that it is possible to get many different kinds of outcomes depending on the degree of competition in each segment of the market. The next section analyzes how these outcomes can change over time.

[20] If our model was extended by adding random price improvement from the quoted bid or ask prices, then you could easily get a full reversal where *EHS* of the third market > *EHS* of the primary market, but the *TTC* of the third market < *TTC* of the primary market.

## 3. Using total trading cost to analyze the evolution of purchasing over time

So far we have stressed that our new measure the total trading cost captures the entire cost of trading, both the effective half spread and the commission. In this section, we show that the total trading cost can be used to analyze the evolution of purchasing over time. Over the past twenty years, the purchase of order flow has evolved through three stages:

- prior to the early eighties, it had not been invented,
- during the mid-eighties, Madoff had innovated it and had a monopoly, and
- from the late eighties on, more and more competitors entered into the purchase of order flow. We analyze this evolution by tracking our model through three stages: (1) nonexistent purchasing, (2) monopolistic purchasing, and (3) competitive purchasing. As a result of this analysis, we are able to resolve a conflict in the empirical literature.

### *3.1. The overview of purchasing equilibria*

In this subsection we examine four different versions of the purchase of order flow model. The four cases are based on different combinations of agents — whether the primary dealers are competitive or monopolistic and whether the purchasers are competitive or monopolistic. Proposition 3 summarizes the these four cases and the types of resulting pure-strategy equilibria.[21]

*Proposition 3. Four different agent combinations of the purchase of order flow model and the resulting pure-strategy equilibria are summarized below*:

| Agent combinations | | | The resulting equilibria | |
|---|---|---|---|---|
| Primary dealers | Purchaser | Subscript | Type of equilibria | Subscript |
| Competitive | Monopolistic | cm | Separating monopolistic | sm |
| Competitive | Competitive | cc | Separating competitive | sc |
| Monopolistic | Monopolistic | mm | Separating monopolistic | sm |
| | | | Pooling monopolistic | pm |
| Monopolistic | Competitive | mc | Separating competitive | sc |

[21] For the monopolistic primary dealer and monopolistic purchaser case, there is a mixed-strategy equilibrium as well. The primary dealer moves first and can randomize between setting an ask price equal to the separating monopolistic ask $a_{sm}$ or above the separating monopolistic ask $a_{sm}$. The type of the equilibrium is determined by the realized ask price. If the realized ask price is equal to the separating monopolistic ask $a_{sm}$, then the actions of the other agents generates a separating monopolistic equilibrium. If the realized ask price is above the separating monopolistic ask $a_{sm}$, then the actions of the other agents generates a pooling monopolistic equilibrium.

To illustrate how the table in Proposition 3 is interpreted, consider the first row. We see the case of competitive primary dealers and monopolistic purchasing (subscript cm) results in a Separating Monopolistic equilibrium (subscript sm) as described in Proposition 1. In one case, multiple equilibria are obtained. From row 3, we see the case of a monopolistic primary dealer and monopolistic purchasing results in either a Separating Monopolistic equilibrium (sm) as described in Proposition 1 or Pooling Monopolistic equilibria (pm) as described in Lemma 1 below.

*Lemma 1. When there is a monopolistic primary dealer and a monopolistic purchaser* (*mm*), *there is a range of possible pooling monopolistic* (*pm*) *equilibria which may result. They are characterized by*

$$p_{\mathrm{mm}} = \varepsilon, \tag{28}$$

$$s_{\mathrm{mm}} = both, \tag{29}$$

$$c_{\mathrm{mm}} = \varepsilon, \tag{30}$$

$$a_{\mathrm{mm}} \in (a_{\mathrm{sm}}, v_{\mathrm{H}}], \tag{31}$$

$$x_{\mathrm{mm}} = \begin{cases} 1 & when\ v \geqslant q + TTC_{\mathrm{mm}}, \\ 0 & otherwise, \end{cases} \tag{32}$$

$$q = \mathrm{E}[v], \tag{33}$$

$$EHS_{\mathrm{mm}} = a_{\mathrm{mm}} - q, \tag{34}$$

$$TTC_{\mathrm{mm}} = EHS_{\mathrm{mm}} + c_{\mathrm{mm}}, \tag{35}$$

$$\mathrm{E}[RHS_{\mathrm{mm}}] = 0, \tag{36}$$

*where* $a_{\mathrm{sm}}$ *is defined in Proposition* 1.

Unfortunately for the monopolistic primary dealer, the monopolist purchaser moves second and can undercut any ask $a_{\mathrm{mm}}$ that would yield positive expected profits. Thus, we describe multiple equilibria which yield zero expected profits for the primary dealer. There is a range of equilibria, which we call the *pooling monopolist equilibria*, in which the monopolist primary dealer attempts to earn positive rents by setting his ask $a_{\mathrm{mm}}$ greater than the *separating monopolistic* ask (as defined in Proposition 1). However, the monopolistic purchaser is able to offer an arbitrarily small payment $p_{\mathrm{mm}} = \varepsilon$, take both the professional and nonprofessional buys, and leave the monopolist primary dealer with nothing.

### 3.2. The no payment for order flow benchmarks

In this subsection, we develop benchmark cases based on the absence of purchasing. First, we analyze the *competitive* primary dealers and *no* payment for order flow case and use the double subscript cn. In the absence of payment for order flow, primary dealers trade the entire pool of orders. Competitive primary dealers earn zero expected profits and the ask price equals the mean of the terminal value conditional on both types of buy orders. Second, we analyze a monopolistic primary dealer. The monopolist extracts the maximum feasible rent by setting the ask price equal to the upper bound of the terminal value.

*Lemma 2. Competitive primary dealers and no payment for order flow* (*cn*) *yield a pooling competitive* (*pc*) *equilibrium*, *which is characterized by*

$$p_{\mathrm{cn}} = 0, \tag{37}$$

$$s_{\mathrm{cn}} = \textit{none}, \tag{38}$$

$$c_{\mathrm{cn}} = 0, \tag{39}$$

$$\begin{aligned} a_{\mathrm{cn}} &= \mathrm{E}[v \mid t \in \textit{both},\ TTC_{\mathrm{cn}}] \\ &= w_{\mathrm{cn}}\mathrm{E}[v \mid v \geqslant q + TTC_{\mathrm{cn}}] + (1 - w_{\mathrm{cn}})\mathrm{E}[v], \end{aligned} \tag{40}$$

$$x_{\mathrm{cn}} = \begin{cases} 1 & \textit{when } v \geqslant q + TTC_{\mathrm{cn}} \\ 0 & \textit{otherwise}, \end{cases} \tag{41}$$

$$q = \mathrm{E}[v] \tag{42}$$

$$EHS_{\mathrm{cn}} = a_{\mathrm{cn}} - q \tag{43}$$

$$TTC_{\mathrm{cn}} = EHS_{\mathrm{cn}} \equiv TTC_{\mathrm{pc}} \equiv TTC_{\mathrm{comp}} \tag{44}$$

$$\mathrm{E}[RHS_{\mathrm{cn}}] = 0 \tag{45}$$

*where* $w_{\mathrm{cn}}$ *is given in the appendix*

- *A monopolistic primary dealer and no payment for order flow* (mn) *yield a pooling monopolistic* (pm) *equilibrium*, *which is characterized by*

$$a_{\mathrm{mn}} = v_{\mathrm{H}}, \tag{46}$$

$$EHS_{\mathrm{mn}} = v_{\mathrm{H}} - q, \tag{47}$$

$$TTC_{\mathrm{mn}} = EHS_{\mathrm{mn}}, \tag{48}$$

$$\mathrm{E}[RHS_{\mathrm{mn}}] = v_{\mathrm{H}} - a_{\mathrm{cn}}. \tag{49}$$

Lemma 2 provides a key benchmark: the competitive total trading cost $TTC_{\text{comp}}$. Following the chain of equalities, the *competitive TTC* is ultimately equal to the mean of the terminal value conditional on both types of buy orders minus $q$ and this is the *TTC* in a *pooling competitive* equilibrium. In the *pooling monopolistic* equilibrium, the ask is set equal to the maximum price ($a_{\text{mn}} = v_{\text{H}}$). This is optimal because it maximizes the profits gained on uninformed buys and simultaneously reduces to zero losses to informed traders. With a commission of zero, the monopolistic total trading cost $TTC_{\text{mn}}$ is also equal to $v_{\text{H}} - q$. The MN equilibrium is the only equilibrium that we analyze in which the primary dealers' expected profit (equal to the expected value of the realized half spread E[$RHS$]) is not zero.

Using our new benchmarks, we obtain the following proposition.

*Proposition 4. The separating competitive TTC equals the competitive TTC*:

$$TTC_{\text{sc}} = TTC_{\text{comp}}. \tag{50}$$

The proposition shows that the total trading cost in a market with competitive internalizers or competitive purchasers equals the total trading cost in a market where all orders are executed by competitive dealers. This should serve as a caution for both empirical and theoretical research comparing the quality of execution across markets. Empirically, it is important use a measure like *TTC* that incorporates commission costs. *TTC* calculates the *total cost* to the customer in order to compare across markets. It is feasible that deep-discount commissions and poor market execution *may* yield superior results to significant commissions and excellent market execution. Theoretically, it is important to include commissions when comparing trading costs or other equilibrium properties across different market designs.

We have now built a substantial the repertoire of equilibria. The next subsection will compare these different types of equilibria.

### 3.3. A numerical calibration

In this subsection, we provide a numerical calibration of the model in order to demonstrate the properties of the equilibria. For Sections 3.3 and 3.4 only, we assume that the terminal value $v$ is uniformly distributed over the interval $[v_{\text{L}}, v_{\text{H}}]$. This simplifies the calculations and permits analytic solutions for some of the equilibria (see the appendix). To calibrate the model, we choose the US Air Group, Inc. stock (symbol U) over the period October 1, 1990 to December 21, 1990. This allows us to match the estimates of Easley et al. (1996) for US Air over this period. They estimate the percentage of informed trades in the primary market (NYSE) = 19.81 percent and the percentage of informed trades in a market were most order flow is purchased (CSE) = 10.26 percent. We numerically backsolve (given the other parameters) for the values of $\alpha^{pro} = 0.2134$ and

$\alpha^{non} = 0.1116$, which generate the same two percentages of informed trading in our model. Interpreting our model as applying over a trading day, we set $v_H = \$15.56$ and $v_L = \$14.81$, which are the average daily high price and average daily low price on US Air stock over the same period. Assuming that half of the uninformed traders are buyers and half are sellers, we set $\Pr(u = 1) = 0.5$. In a separating equilibrium, all of the trades in the primary market are nonprofessional, so we set $\beta = 0.65$ based on the Cochrane (1993) estimate that 65 percent of trades for 100 to 2099 shares are traded on the primary market (NYSE).

The separating competitive equilibrium of our calibrated model yields an *EHS* in the primary market of 4.26 cents, which is equal to the *EHS* in the third market. However, the *TTC* differs between the two markets. The *TTC* in the primary market is 4.26 cents since there is no payment for order flow and thus, no commission rebate. The *TTC* in the third market is 3.54 cents due to the commission rebate. Hence, the *TTC* tells the real story, whereas the *EHS* does not.

Going a step further, our calibrated commission rebate is $-c = TTC - EHS = 0.72$ cents. We can solve for the implied purchase of order flow amount $p = -c/\Pr(t \in non \mid buy, TTC_{sc}) = 2.05$ cents. This level of payment for order flow is consistent with the actual level of one to two cents.

Fig. 3 graphs two side-by-side bars, the *EHS* is the left bar and the *TTC* is the right bar, for each of the four types of equilibria: (1) separating monopolistic

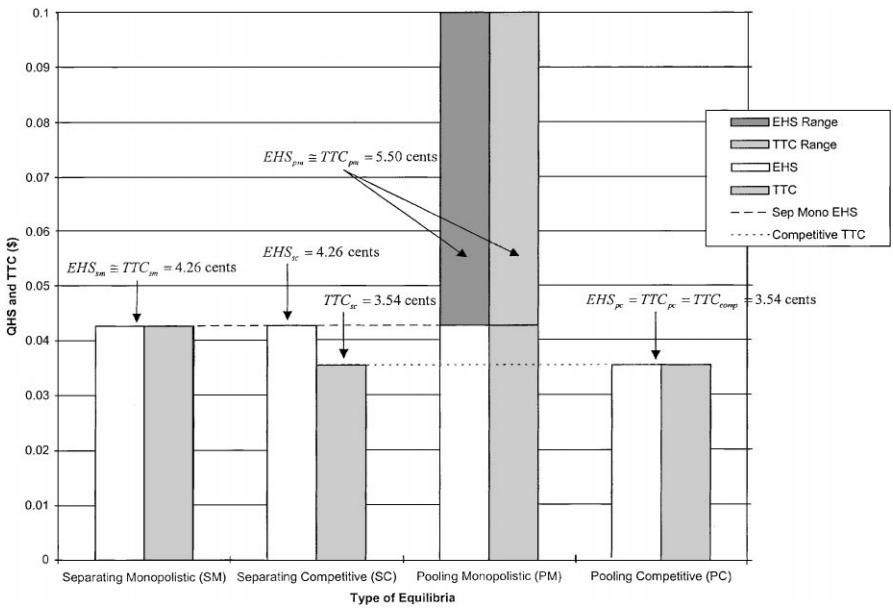



Fig. 3. QHS and TTC by type of equilibria.

(sm), (2) separating competitive (sc), (3) pooling monopolistic (pm), and (4) pooling competitive (pc). The graph illustrates the key differences between these equilibria. First, we notice that the two monopolistic equilibria (sm and pm) yield a *TTC* which (arbitrarily closely) equals the *EHS*. In the sm case, $EHS_{\mathrm{sm}} \cong TTC_{\mathrm{sm}} = 4.26$ cents. As an example of the pm case, $EHS_{\mathrm{pm}} \cong TTC_{\mathrm{pm}} = 5.50$ cents. Both $EHS_{\mathrm{pm}}$ and $TTC_{\mathrm{pm}}$ are elements of a range (4.26 cents, 39.50 cents] (i.e., a range running from $a_{\mathrm{sm}} - q$ up to $v_{\mathrm{H}} - q$). Secondly, we notice that the separating competitive equilibrium yields a *TTC* which is lower than its *EHS*. Specifically, $TTC_{\mathrm{sc}} = 3.54$ cents is less than the $EHS_{\mathrm{sc}} = 4.26$ cents. Thirdly, consistent with Proposition 4, we notice that the *separating competitive TTC* equals the *pooling competitive TTC* equals the *competitive TTC* ($TTC_{\mathrm{sc}} = TTC_{\mathrm{pc}} = TTC_{\mathrm{comp}} = 3.54$ cents).

Finally, we notice that the lower bound of the *pooling* monopolistic *EHS* is the separating monopolistic *EHS* (4.26 cents = large dashes). The intuition for this is based on the optimal strategy of the purchaser(s). *Above* the lower bound it is optimal to purchase both professional and nonprofessional orders (pooling), but *at* the lower bound professional orders generate losses and it is optimal to switch to purchasing nonprofessional orders only (separating). Prices *below* the lower bound are not equilibria, because at those prices purchaser(s) would prefer to purchase nonprofessional orders only, but primary dealers would lose money on professional orders only.

### *3.4. Evolution of purchasing over time and resolving a conflict*

This subsection analyzes the evolution of purchasing over time and uses this analysis to resolve a conflict in the empirical literature. We follow the evolution of purchasing through three stages: (1) nonexistent purchasing, (2) monopolistic purchasing, and (3) competitive purchasing. Our theory permits *three* possible paths depending on whether the primary dealers are competitive or monopolistic and on whether the equilibrium in stage two is pooling or separating. For each path, we trace the patterns over time of the *TTC* and probability of informed trading on and away from the primary market.

We start with a numerical example in Fig. 4 to illustrate the three paths. This example is based on the same parameters as the numerical calibration in Section 3.3. Following the numerical example, we develop a proposition which yields the same qualitative features under the minimal distributional assumptions of Section 2.

Fig. 4 graphs the *TTC* on the $y$-axis for each of the three paths against the evolution of purchasing over time on the $x$-axis going from stage one, nonexistent purchasing (left side), to stage two, monopolistic purchasing (center), to stage three, competitive purchasing (right side). Table 3 summarizes how *TTC* changes during the transitions along the three paths are:

To elaborate, Path 1 is based on a monopolistic primary dealer. It starts at the monopolistic *TTC* of 39.50 cents when there is no purchaser, then decreases to

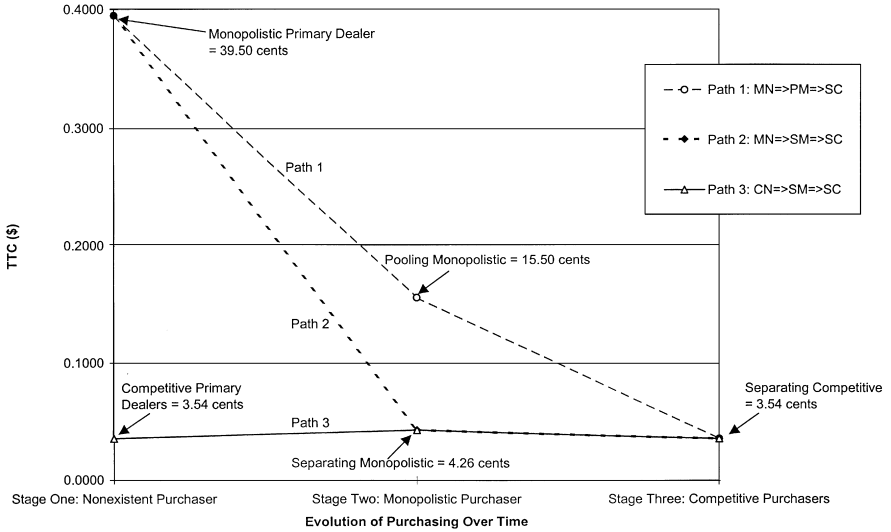



Fig. 4. TTC for each path by the evolution of purchasing over time.

a pooling monopolistic *TTC* of 15.50 cents when the monopolist purchaser enters, and then decreases further to the competitive *TTC* of 3.54 cents when competing purchasers enter. Path 2 is also based on a monopolistic primary dealer. It starts at the monopolistic *TTC* of 39.50 cents when there is no purchaser, then decreases to a separating monopolistic *TTC* of 4.26 cents when the monopolist purchaser enters, and then decreases further to the competitive *TTC* of 3.54 cents when competing purchasers enter. Path 3 is based on competitive primary dealers. It starts low at the Competitive *TTC* of 3.54 cents when there is no purchaser, then *increases* to a separating monopolistic *TTC* of 4.26 cents when the monopolist purchaser enters, and then decreases to the competitive *TTC* of 3.54 cents when competing purchasers enter.

For the rest of the paper, we drop the assumption of uniform distributions and generalize to the original distributional assumptions specified in Section 2. In the proposition below we are also interested in characterizing the probability of informed trading in different venues under the *separating* equilibria. Consider the *separating competitive* equilibrium where the third-market dealer purchases nonprofessional buys and the primary dealer executes professional buys. The probability of informed trading is $w_{sc}^{non}$ in the third market and $w_{sc}^{pro}$ in the primary market, where these weights on informed trading where calculated in Proposition 2. This leads to the next proposition.

*Proposition* 5. *Under the original distributional assumptions specified in Section* 2, *there are three time paths that are consistent with our theory*:

| | Transition from stage one (no purch.) to stage two (mono purch.) | Nature of stage two equilibrium | | Transition from stage two (mono.purch.) to stage three (comp purch.) | Nature of stage three equilibrium | |
|---|---|---|---|---|---|---|
| | | Orders that get purchased | Prob. of informed trading | | Orders that get purchased | Prob. of informed trading |
| Path 1: MN ⇒ PM ⇒ SC | *TTC* decreases | All orders | Not applicable | *TTC* decreases | Nonprof. | Third market is lower |
| Path 2: MN ⇒ SM ⇒ SC | *TTC* decreases | Nonprof. | Third market is lower | *TTC* decreases | Nonprof. | Third market is lower |
| Path 3: CN ⇒ SM ⇒ SC | *TTC* increases | Nonprof. | Third market is lower | *TTC* decreases | Nonprof. | Third market is lower |

Table 3
TTC changes during the transitions along three paths

| | Transition from stage one (no purch.) to stage two (mono purch.) | Transition from stage two (mono purch.) to stage three (comp purch.) |
|---|---|---|
| Path 1: MN ⇒ PM ⇒ SC (long dashes) | *TTC* decreases | *TTC* decreases |
| Path 2: MN ⇒ SM ⇒ SC (short dashes) | *TTC* decreases | *TTC* decreases |
| Path 3: CN ⇒ SM ⇒ SC (solid line) | *TTC* increases | *TTC* decreases |

This proposition serves as the basis for resolving a conflict in the empirical literature on the issue of 'cream skimming' vs. cost competition. Recent empirical studies claim opposite results on this issue. We distinguish two separate dimensions: (1) price impact in the primary market and (2) differences in informativeness between the primary market and the third market.

On the first dimension, Battalio (1997), combined with deep-discount commissions by on-line brokers that accept payment for order flow, shows that the *TTC* drops when Madoff begins purchasing order flow, which he interprets as consistent with cost competition. On the second dimension, Easley et al. (1996) find that the probability of informed trading is lower on the CSE than the NYSE and interpret this evidence to be consistent with cream-skimming.

We resolve this conflict by showing that our theoretical model can generate *both* results simultaneously. Specifically, Path 2 can generate: (1) a *drop* in the *TTC* and (2) separating equilibria with a *lower* probability of informed trading in the third market than the primary market. Further, the combined empirical evidence is consistent with Path 2, inconsistent with Path 1 (because purchasers do not purchase all of the order flow), and inconsistent with Path 3 (because the *TTC* did not increase when Madoff entered). We conclude that the combined evidence of Battalio (1997) and of Easley et al. (1996) are consistent with Path 2 of our theory and are therefore consistent with each other. Further, we note the implications of the empirically-supported Path 2 are that:

- prior to the purchase of order flow, the NYSE was extracting some amount of monopolistic rents,
- upon the entry of Madoff, the level of rents was reduced, and
- upon the entry of competing purchasers, the level of rents was reduced further.

## 4. Internalization with internally-verifiable characteristics

In this section, we develop an additional explanation for why decimal trading does *not* eliminate internalization. Specifically, we analyze what happens when brokers are able to use their relationships and personal contacts with customers to sort traders and their order flows using internally-verifiable characteristics (characteristics derived from the broker/dealer relationship which are either subjective and/or prohibitively expensive for a third-party to verify). We show that brokers can profit by internalizing some trades and not others. This creates a key distinction between internalization and the purchase of order flow. Brokers have the advantage of being able to selectively internalize orders using *both* externally-verifiable and internally-verifiable characteristics. By contrast, third-market dealers are limited to selectively purchasing orders using *externally-verifiable characteristics only*, since they are formally contracting with the brokers.

We make one modification to the internalization model in Section 2.3 to capture the (realistic) case in which brokers can sort orders based on internally-verifiable characteristics. It is easy to imagine dozens characteristics that might be useful in forecasting whether the trader is informed or not. For example, customer occupation, degree of customer investment sophistication, outside customers vs. competing market makers, etc. We model each broker as being able to subjectively judge if a trader is 'sophisticated' or 'naïve', where sophisticated traders are more likely to be informed than naïve traders. This internally-verifiable distinction is in addition to the externally-verifiable distinction between professional vs. nonprofessional traders.

Fig. 5 shows the modified arrival process for traders. Let $\gamma$ be the probability that a professional trader is sophisticated, $1-\gamma$ be the probability that a professional trader is naïve, $\delta$ be the probability that a nonprofessional trader is sophisticated, and $1-\delta$ be the probability that a nonprofessional trader is naïve.

We adopt the convention that professional traders are more likely to be informed than nonprofessional traders and sophisticated traders are more likely to be informed than naïve traders. Hence, we get the following orderings $\alpha^{ps} > \alpha^{pn} > \alpha^{nn}$ and $\alpha^{ps} > \alpha^{ns} > \alpha^{nn}$. The key implication of this convention is that Professional Sophisticated traders ($t = \text{ps}$) have the highest probability of being informed and they will drive the resulting equilibria.

Fig. 6 provides a timeline illustration of the expanded model on the (trader) buy side based on brokers who condition on both internally-verifiable and externally-verifiable characteristics when deciding which orders to internalize.

To solve the model we start at Step 5, the broker's decision on whether to send the received market order to the primary dealer or internalize it. As with the previous internalization model in Section 2.3, the broker's optimal policy is to internalize an order when the expected terminal value conditional on that type

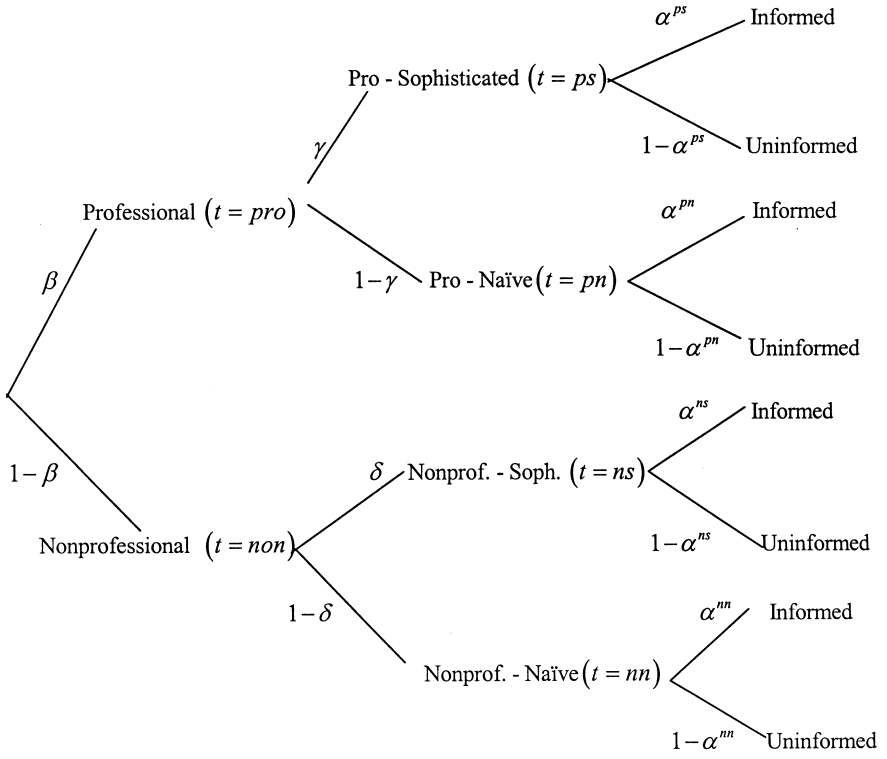


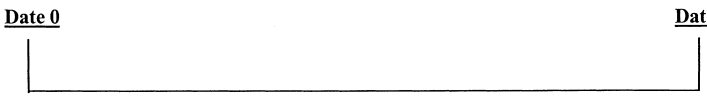

Fig. 5. Modified arrival process for traders.

of order is less than the ask $a$ and pass it through to the primary dealer otherwise.

For Step 3, we add notation for the broker's internalization decision. In the modified Step 3, the $b$th broker anticipates making a profit margin of $\pi^b$ on all orders in the internalization set $i^b$ and chooses $c^b$ and $i^b$ to maximize expected profits

$$\underset{c^b, i^b}{\text{Max}} \; (f^b)\,\mathrm{E}[(\pi^b + c^b) \mid t \in i^b] + (1 - f^b)\mathrm{E}[c^b \mid t \notin i^b]. \tag{51}$$

In equilibrium, there will be a single, competitive commission $c$, probability $f$ of receiving a buy order in the competitive internalization set $i$, and competitive profit margin $\pi$. As was the case when order flow was purchased, internalizing brokers earn profits which are competed away in the form of negative commissions. Setting the broker's expected profits to zero and solving for $c$, we obtain $c = -f\pi$.




Step 1: The $i$th primary dealer quotes an ask $a^i$ for all buy orders.

Step 3: The $b$th broker offers a commission $c^b$ for all buy orders.

Step 4: A trader arrives, chooses to submit a buy or do nothing, and chooses a broker to handle the order. A trader of type $t$ $\left(t \in \{ps, pn, ns, nn\}\right)$ arrives. Then:

- With probability $\alpha^t$, the trader is informed and chooses whether to buy $\left(x^t = 1\right)$ or not $\left(x^t = 0\right)$.
- With probability $1-\alpha^t$, the trader is uninformed and a random number of shares $u^t \in \{0,1\}$ is drawn.

Step 5: The broker receiving the market buy:

- observes the trader type
- updates his conditional expectation of the terminal value, and
- chooses whether to send it to the primary dealer or internalize it.

Step 6: The market buy is executed.

The security's terminal value $v$ is realized

Fig. 6. Expanded internalization model with internally-verifiable characteristics.

Table 4
New choice variable for the expanded internalization model with internally-verifiable char[a]

| Class of agents | Choice variables | Space of the choice variable |
|---|---|---|
| Brokers | $i^b$ = Internalization set of the $b$th broker | $\{none, ps, pn, ns, nn, \ldots, all\}$ |

[a] $i^b = ps$ means internalize professional sophisticated.

From the conventions that $\alpha^{ps} > \alpha^{pn} > \alpha^{nn}$ and $\alpha^{ps} > \alpha^{ns} > \alpha^{nn}$, it follows that the broker's conditional expectations follow the same orderings $\mathrm{E}[v \mid ps] > \mathrm{E}[v \mid pn] > \mathrm{E}[v \mid nn]$ and $\mathrm{E}[v \mid ps] > \mathrm{E}[v \mid ns] > \mathrm{E}[v \mid nn]$. Hence, if the ask price is driven by the class of *professional sophisticated* traders, then broker's optimal strategy is to internalize the relatively less-informed orders ($i^b = \{pn, ns, nn\}$).

Step 1 follows the internalization case in Section 2.3. We obtain the following proposition.

*Proposition 6. When there are competitive primary dealers and internalization by competitive brokers using both internally-verifiable and externally-verifiable characteristics (CIB), the resulting equilibrium is super separating competitive (SSC) as characterized by*

$$i_{cib} = \{pn, ns, nn\}, \tag{52}$$

$$c_{cib} = -\Pr(t \notin ps \mid buy, TTC_{cib})(a_{cib} - m_{cib}), \tag{53}$$

$$\begin{aligned} a_{cib} &= \mathrm{E}[v \mid t \in ps, TTC_{cib}] \\ &= w_{cib}^{ps}\mathrm{E}[v \mid v \geqslant q + TTC_{cib}] + (1 - w_{cib}^{ps})E[v], \end{aligned} \tag{54}$$

$$\begin{aligned} m_{cib} &= \mathrm{E}[v \mid t \notin ps, TTC_{cib}] \\ &= w_{cib}^{nops}\mathrm{E}[v \mid v \geqslant q + TTC_{cib}] + (1 - w_{cib}^{nops})E[v], \end{aligned} \tag{55}$$

$$x_{cib} = \begin{cases} 1 & when\ v \geqslant q + TTC_{cib}, \\ 0 & otherwise, \end{cases} \tag{56}$$

$$q = \mathrm{E}[v], \tag{57}$$

$$EHS_{\mathrm{cn}} = a_{cib} - q, \tag{58}$$

$$TTC_{\mathrm{cn}} = EHS_{\mathrm{cn}} + c_{cib} \equiv TTC_{ssc}, \tag{59}$$

$$\mathrm{E}[RHS_{\mathrm{cn}}] = 0, \tag{60}$$

*where* $w_{cib}^{ps}$, $w_{cib}^{no}$, *and* $\Pr(t \notin ps \mid buy, TTC_{cib})$ *are given in the appendix.*

- *When there are monopolistic primary dealers and internalization by competitive brokers using both internally-verifiable and externally-verifiable characteristics (MIB), the resulting equilibrium is also super separating competitive (SSC).*
- *Further, the super separating competitive TTC equals the competitive TTC*

$$TTC_{ssc} = TTC_{comp}. \tag{61}$$

This proposition demonstrates that the optimal strategy for brokers is to pass through to the primary dealers the class of orders with the highest probability of being informed (professional sophisticated) and internalize the rest. This section demonstrates that broker internalization using internally-verifiable characteristics (in addition to externally-verifiable characteristics) is an additional explanation of internalization – one that does not depend on the tick size.

## 5. Revisiting the performance of different trading venues

In this section, we revisit five empirical studies of execution performance by trading venue. All five studies use *EHS* as the basis of comparison. Each of these studies compares the *EHS* on the NYSE, which does not purchase order flow, to the *EHS* on the regional exchanges and/or the Third Market, trading venues on which order flow is purchased and/or internalized. We revisit 13 comparisons, five of which are NYSE versus NASD's Third Market and eight of which are NYSE versus regional exchanges. The studies report that the NYSE had a lower *EHS* in 11 of the 13 comparisons, whereas the non-NYSE venue had a lower *EHS* in only 2 comparisons.

This paper has introduced the concept of total trading cost (*TTC*) and we would prefer to directly estimate *TTC* to do a new comparison. However, to directly estimate *TTC* we would need data on: (1) commissions by broker and (2) orders submitted by broker. We do not have this data and so this task is left for future research.

However, we can estimate a closely related concept even without the brokerage data. Define the *TTC-Equalizing Passthrough Percentage* as the percentage of the payment for order flow that the broker must passthrough to the trader in order to equalize the *TTC* of both venues. Formally, define

$$TTC\text{–}Equalizing\ Passthrough\ Percentage = \frac{EHS_{\text{Payment market}} - EHS_{\text{No payment market}}}{Payment\ amount}. \tag{62}$$

The idea is to determine what percentage of payment amount would have to be passed through to the trader in order to completely offset the difference in *EHS* between the two venues and thus equalize the *TTC* of the two venues.

Table 5 reports the *TTC-Equivalent Passthrough Percentage* corresponding to 13 *EHS* comparisons in five studies. For each study, it shows the trade size, time period, and spread width involved in each *EHS* comparison.[22] Actual payments for order flow during this time period ranged from one to two cents per share. We use two cents per share as our benchmark payment amount to calculate the *TTC-Equivalent Passthrough Percentage*.

[22] With the exception of Battalio (1997), all estimates of the *EHS* used to determine the break-even passthrough rates are generated from relatively small trades. *EHS* comparisons are typically reported conditional on the spread width, since the distribution of *EHS* is discrete and is a function of the spread. Assuming that firms which internalized order flow could have sold for a payment, the internalization profit per share likely exceeds the benchmark payment for order flow.

Table 5
*TTC*-equalizing passthrough percentage

| Academic study | Trade size | Time period | Spread width | *TTC*-equalizing passthrough percentage = percentage of the payment for order flow[a] that the broker must passthrough to the trader in order to equalize the *TTC* of both venues | | |
|---|---|---|---|---|---|---|
| | | | | NYSE[b] vs. third market[c] | NYSE vs. best regional[d] | NYSE vs. worst regional[e] |
| Lee (1993) | 100–499 | 1991 | All[j] | 50%[f] | 35% | 255% |
| Battalio (1997) | All | 1988–90 | 1/4[j] | 95% | n.a.[g] | n.a. |
| BGJ (1997)[h] | 100–1299 | 1994–5 | 1/4[j] | 100% | 50% | 75% |
| SEC (1997) | 101–200 | 1996 | 1/8[k] | n.a. | 35% | 70% |
| | | | 1/4[k] | n.a. | − 20% | 220% |
| BHJ (2000)[i] | 100–499 | 1999 | 1/16[k] | 35% | n.a. | n.a. |
| | | | 1/8[k] | − 80% | n.a. | n.a. |

[a]Based on a two cent per share payment for order flow.
[b]NYSE refers to the New York Stock Exchange.
[c]3rd Market refers to the NASD's Third Market.
[d]Best regional refers to the regional exchange with the best execution prices.
[e]Worst regional refers to the regional exchange with the worst execution prices.
[f]Percent of two cents per share payment that must be passed on by brokers via lower commissions to ensure investors are not harmed by the routing of orders in NYSE-listed securities to the NASD's 3rd Market rather than to the NYSE.
[g]Statistic not reported in paper.
[h]Battalio et al. (1997).
[i]Battalio et al. (2000).
[j]Statistics computed versus execution time spread.
[k]Statistics computed versus order receipt time spread.

Turning to the results in the last three columns of Table 5, the estimates of Lee (1993) for NYSE vs. Third Market imply a *TTC*-Equivalent Passthrough Percentage of 50 percent. This implies that, on average, 50 percent of a two cent per share payment needs to flow through to investors to give them the same *TTC* in both the Third Market and the NYSE during Lee's sample period. As can be seen from Table 5, there are two comparisons in which the *TTC-Equivalent Passthrough Percentage* is greater than 100% (**boldface**) and thus the NYSE is superior independent of the actual passthrough percentage. There are two comparisons in which the *TTC-Equivalent Passthrough Percentage* is negative (gray background) and thus the non-NYSE trading venue offers lower *TTC* even with zero passthrough percentage. Nine of the 13 comparisons are intermediate cases. Which way the *TTC* ranking goes in these cases depends the degree of competition in the brokerage industry and what passthrough percentage actually takes place in the real-world. Suppose the degree of competition is sufficient

for more than 50 percent of the payment for order flow to actually passthrough to the trader. In this case, five of the 11 comparisons that initially favored the NYSE (involving four of the five studies) would be reversed. That is, the majority of the comparisons (7 out of 13) the *TTC* would be lower on the non-NYSE venue than the NYSE.

In this section we have shown that many of the performance orderings made on the basis of *EHS* can be reversed when made on the basis of *TTC* if brokers passthrough a significant amount of the payment for order flow to the trader. The lack of discount brokers who do not sell or internalize order flow suggests that a meaningful amount of payment for order flow is passed on to traders in the form of lower commissions. The variation in the *TTC-Equivalent Passthrough Percentage* within and across samples suggests it is dangerous to ignore payment for order flow and internalization. We suggest that future performance studies calculate the *TTC-Equivalent Passthrough Percentage* in order to investigate the robustness their conclusions to the impact of payment for order flow and internalization on brokerage commissions.

## 6. Conclusion

We make four contributions to the literature. First, we show that *externally-verifiable characteristics* of traders and/or orders allow profitable purchasing or profitable internalization, even when prices are permitted on the real line and primary dealers are competitive. Second, we define total trading cost and use this concept to reconcile the otherwise contradictory results of Easley et al. (1996) and Battalio (1997). We show that it is possible for *TTC* to fall when purchasers enter the market and for the diverted order flow to be less informed than the orders sent to the primary market. Third, we show that brokers can use their direct relationships with customers to *internally-verify characteristics* of traders and/or orders. This additional information allows brokers to perform additional sorting of the order flow, above and beyond what purchasers can do, and engage in further profitable internalization. Fourth, we revisit the empirical literature on performance by trading venue to determine the sensitivity of their conclusions to the impact of payment for order flow and internalization on brokerage commissions. We show that many of the performance rankings would be reversed depending on the degree of competition of the brokerage industry. We suggest that future papers on performance by venue calculate the *TTC-Equalizing Passthrough Percentage* as a robustness check.

It would be interesting to expand our model to include fixed costs of setting up an internalization operation, costs of contracting for order flow, and heterogeneous brokerage firms (full service vs. discount). This framework would rationalize the co-existence of internalization and purchase of order flow and allow further comparisons of the two practices. This is left for future research.

## For Further Reading

The following reference is also of interest to the reader: US Securities and Exchange Commission, 1997.

## Appendix A

*Proof of Proposition 1.* The expressions for the weight and probability are:

$$w_{\mathrm{cm}} \equiv \frac{\alpha^{pro}\Pr[v \geqslant q + TTC_{\mathrm{cm}}]}{\alpha^{pro}\Pr[v \geqslant q + TTC_{\mathrm{cm}}] + (1-\alpha^{pro})\Pr[u=1]}, \tag{A.1}$$

$$\Pr(t \in non \,|\, buy, TTC_{\mathrm{scm}})$$

$$\equiv \frac{(1-\beta)(\alpha^{non}\Pr[v \geqslant q + TTC_{\mathrm{cm}}] + (1-\alpha^{non})[\Pr(u=1)])}{(1-\beta)(\alpha^{non}\Pr[v \geqslant q + TTC_{\mathrm{cm}}] + (1-\alpha^{non})[\Pr(u=1)]) + (\beta)(\alpha^{pro}\Pr[v \geqslant q + TTC_{\mathrm{cm}}] + (1-\alpha^{pro})[\Pr(u=1)])}. \tag{A.2}$$

Steps 5, 4, and 3 are in the text.

*Step 2*: Eq. (6) can be split into two parts

$$\max_{p_m, q_m} \Pr(t \in non \,|\, buy, TTC_{\mathrm{cm}})\{(a_{\mathrm{cm}} - p_{\mathrm{cm}}) - \mathrm{E}[v]\} \tag{A.3}$$

$$+ \begin{cases} \Pr(t \in pro \,|\, buy, TTC_{\mathrm{cm}})\{(a_{\mathrm{cm}} - p_{\mathrm{cm}}) - E[v \,|\, t = pro, TTC_{\mathrm{cm}}]\} & \text{when } s_{\mathrm{cm}} = \text{both}, \\ 0 & \text{otherwise.} \end{cases} \tag{A.4}$$

Hence, the optimal policy is

$$p_{\mathrm{cm}} = \varepsilon \quad s_{\mathrm{cm}}$$

$$= \begin{cases} both & \text{when the profit margin } a_{\mathrm{cm}} - \varepsilon - \mathrm{E}[v \,|\, t = pro, TTC_{\mathrm{cm}}] > 0, \\ non & \text{otherwise.} \end{cases} \tag{A.5}$$

where $\varepsilon > 0$ is arbitrarily small.

*Step 1*: From Step 2, the only order flow which the primary dealer might receive are professional orders. Hence, equation (8) the zero expected profit condition can be decomposed into two terms

$$(\alpha^{pro}\Pr[v \geqslant q + TTC_{\mathrm{cm}}])(a_{\mathrm{cm}} - \mathrm{E}[v \,|\, v \geqslant q + TTC_{\mathrm{cm}}])$$
$$+ ((1-\alpha^{pro})\Pr[u=1])(a_{\mathrm{cm}} - E[v]) = 0. \tag{A.6}$$

Then solve this equation for the competitive ask $a_{cm}$. Since the primary dealers set the ask to break even on professional orders, the purchaser would loose money by paying a positive amount to buy professional orders. Thus, the optimal contract is $s_{cm} = non$. Substituting this into (3) yields $c_{cm}$, which in turn yields $a_{cm}$ and $x_{cm}$. □

*Proof of Proposition 2.* The expressions for the weights and probability are:

$$w_{ci}^{pro} \equiv \frac{\alpha^{pro}\Pr[v \geqslant q + TTC_{ci}]}{\alpha^{pro}\Pr[v \geqslant q + TTC_{ci}] + (1-\alpha^{pro})\Pr[u=1]}, \tag{A.7}$$

$$w_{ci}^{non} \equiv \frac{\alpha^{non}\Pr[v \geqslant q + TTC_{ci}]}{\alpha^{non}\Pr[v \geqslant q + TTC_{ci}] + (1-\alpha^{non})\Pr[u=1]}, \tag{A.8}$$

$$\Pr(t \in non \,|\, buy, TTC_{ci}) \equiv \frac{(1-\beta)(\alpha^{non}\Pr[v \geqslant q + TTC_{ci}] + (1-\alpha^{non})[\Pr(u=1)]}{(1-\beta)(\alpha^{non}\Pr[v \geqslant q + TTC_{ci}] + (1-\alpha^{non})[\Pr(u=1)]) + (\beta)(\alpha^{pro}\Pr[v \geqslant q + TTC_{ci}] + (1-\alpha^{pro})[\Pr(u=1)])}. \tag{A.9}$$

First, we consider a separating equilibrium in which brokers internalize nonprofessional orders. In this case, the probability $f = \Pr(t = non \,|\, buy, TTC_{ci})$ and is calculated as the ratio of the probability of the two events (non. inf. buy and non. uninf. buy) which yield nonprofessional buys divided by the probability of all four events (non. inf. buy, non. uninf. buy, pro. inf. buy, and pro. uninf. buy). The profit margin is the difference in conditional means $\pi = a_{ci} - m_{ci}$. Substituting into $c = -f\pi$ yields the competitive commission. The competitive ask equals the conditional mean given professional orders $a_{cc} = \mathrm{E}[v \,|\, t \in pro, TTC_{ci}]$ and is calculated as in Step 1 above. The conditional mean $m$ equals the conditional mean given nonprofessional orders $m_{cc} = \mathrm{E}[v \,|\, t \in non, TTC_{ci}]$ and is calculated in a manner analogous to Step 1 above, except that *non* is substituted for *pro* on all subscripts.

Next, consider the possibility of internalizing both types and offering a commission equal to minus the difference in conditional means. This is not an equilibrium because individuals have an incentive to deviate from it. Specifically each individual broker has an incentive to earn positive rents by offering a slightly smaller commission (larger rebate) and internalizing only nonprofessional orders. □

*Proof of Proposition 3. Competitive Monopolistic* (*cm*) *case*. See Proposition 1.

*Competitive competitive* (*CC*) *case*: This case is analogous to Proposition 2. Purchasers purchase nonprofessional orders ($s_{cc} = non$). Given that they are nonprofessional, their conditional mean is $m_{cc} = \mathrm{E}[v \,|\, t \in non, TTC_{cc}]$. Purchasers get an asset worth $m_{cc}$ and get to sell it (execute it) at the ask $a_{cc}$. If they

paid nothing, then their profit margin would be $a_{cc} - m_{cc}$. However, they are competitive and pay out all of their rents. Hence, the competitive payment is $p_{cc} = a_{cc} - m_{cc}$. Similar to Proposition 2, the competitive commission is $c = -fp = -\Pr(t \in non \mid buy, TTC_{cc})(a_{cc} - m_{cc})$. The rest is exactly the same as Proposition 2.

*Monopolistic monopolistic* (*mm*) *case*: The analysis of Steps 5, 4, 3, and 2 are the same as in Section 2.2. Turning to Step 1, the monopolist primary dealer chooses an ask $a_{mm}$ to maximize expected profits

$$\underset{a_{mm}}{\text{Max}}\ \mathrm{E}[a_{mm} - v \mid t \notin s_{mm}, TTC_{mm}] = \underset{a_{mm}}{\text{Max}}\ (a_{mm} - \mathrm{E}[v \mid t \notin s_{mm}, TTC_{mm}]). \tag{A.10}$$

If the monopolistic primary dealer attempts to earn positive expected profits by setting any ask $a_{mm}$ above separating monopolistic ask defined in Proposition 1, then the monopolist purchaser can undercut that price by an arbitrarily small amount and purchase both professional and nonprofessional buys. Thus, we obtain multiple equilibria which yield zero expected profits for the primary dealer. One possibility is a separating monopolistic equilibrium (sm), where the monopolist primary dealer mimics competitive primary dealers by setting the ask $a_{mm}$ equal to the separating monopolistic ask. In this case, it will be unprofitable for the monopolistic purchaser to purchase the professional buys and thus we obtain the identical equilibrium as the separating monopolistic case described in Proposition 1. Alternatively, there is a range of equilibria, which we call the pooling monopolist equilibria, in which the monopolist primary dealer chooses an ask $a_{mm}$ in the interval $(a_{sm}, v_h]$. The monopolistic purchaser offers an arbitrarily small payment $p_{mm} = \varepsilon$ and takes both the professional and nonprofessional buys. This leaves the monopolist primary dealer with nothing.

*Monopolistic competitive* (*mc*) *case*: Steps 5, 4, and 3 are the same as Section 2.2. Turning to Step 2, the competitive purchasers commission returns all of their rents, just as they did in the competitive competitive (cc) case. At a minimum, they will purchase all of the nonprofessional orders. Further, if there are any rents to be made on professional orders, they will purchase them as well. Turning to Step 1, the monopolistic primary dealer would like to set the ask to earn rents on the professional orders. However, any effort to earn positive rents will cause all orders to be purchased and the primary dealer will be left with zero profits. It is not an equilibrium for all competitive purchasers to purchase all of the orders and price them competitively – for the same reasons discussed at the end of the proof of Proposition 2. Essentially, individual purchasers have an incentive to deviate from it. Specifically each individual purchaser has an incentive to earn positive rents by offering a slightly larger payment and contracting for only nonprofessional

orders. Thus, the only equilibrium is separating competitive (sc) as described in Proposition 1. □

*Proof of Lemma 1.* See the monopolistic monopolistic (mm) case of Proposition 3. □

*Proof of Lemma 2.* The expression for the weight is:

$$w_{\mathrm{cn}} \equiv \frac{[\beta\alpha^{pro} + (1-\beta)\alpha^{non}]\Pr[v \geqslant q + TTC_{\mathrm{cn}}]}{[\beta\alpha^{pro} + (1-\beta)\alpha^{non}]\Pr[v \geqslant q + TTC_{\mathrm{cn}}] + [\beta(1-\alpha^{pro}) + (1-\beta)(1-\alpha^{non})][\Pr(u=1)]} \tag{A.11}$$

*Competitive no payment for order flow* (*cn*) *case*: Step 5 is the same as Section 2.2. Obviously with no payment for order flow, the payment is zero, the commission is zero, and the contractual set is none. The competitive ask equals the conditional mean given professional orders $a_{\mathrm{cn}} = \mathrm{E}[v \mid t \in both, TTC_{\mathrm{cn}}]$. This can be calculated as a weighted average of the conditional expectation given an informed trader and the conditional expectation given an uninformed trader. The weight $w_{\mathrm{cn}}$ is calculated as the ratio of the probability of the two events (pro. inf. buy and non. inf. buy) which yield informed buys divided by the probability of all four events (pro. inf. buy, non. inf. buy, pro. uninf. buy, and non. uninf. buy).

*Monopolistic no payment for order flow* (*cn*) *case*: This is a special case of the pooling monopolistic case in Proposition 3. In the absence of payment for order flow, a monopolistic primary dealer will trade the entire pool of orders at a monopolistic no payment for order flow ask $a_{\mathrm{mn}}$ set equal to the maximum price in the pooling monopolistic interval $(a_{\mathrm{sm}}, v_{\mathrm{H}}]$.[23] This is obviously optimal since $a_{\mathrm{mono}} = v_{\mathrm{H}}$ maximizes the profits gained on uninformed buys and simultaneously reduces to zero losses to informed traders. With a commission of zero, the monopolistic $TTC$ ($TTC_{\mathrm{mono}}$) is equal to $v_{\mathrm{H}} - q$. □

*Proof of Proposition 4.* To show that $TTC_{\mathrm{sc}} = TTC_{\mathrm{comp}}$, take the expression for $a_{\mathrm{ci}}$ in (7), subtract $q$, and add the expression for $c_{\mathrm{ci}}$ in (6), then combine and simplify the result and one obtains the expression for $TTC_{\mathrm{comp}}$ in Lemma 1. □

*An analytic solution under the uniform distribution*: Assuming that the terminal value $v$ is uniformly distributed over the interval $[v_{\mathrm{L}}, v_{\mathrm{H}}]$ yields closed form

[23] We assume that the exchange (or regulators) do not permit primary dealers to set an ask price above $v_{\mathrm{H}}$.

formulas for the following probability and conditional expectation in the separating monopolistic equilibrium

$$\Pr(v > q + TTC_{\mathrm{sm}}) = \frac{v_{\mathrm{H}} - (q + TTC_{\mathrm{sm}})}{v_{\mathrm{H}} - v_{\mathrm{L}}}$$

$$\mathrm{E}[v \mid v > q + TTC_{\mathrm{sm}}] = \frac{v_{\mathrm{H}} + (q + TTC_{\mathrm{sm}})}{2} \tag{A.12}$$

and identical closed form formulas (with pc subscripts) for a pooling competitive equilibrium.

Since $\varepsilon$ is arbitrarily small, we analyze the limiting case as $\varepsilon \to 0$. We obtain the following closed form formula for the separating monopolistic ask

$$a_{\mathrm{sm}} = \frac{-d_{\mathrm{sm}} + \sqrt{d_{\mathrm{sm}}^2 - 4b_{\mathrm{sm}}e^{\mathrm{m}}}}{2c_{\mathrm{sm}}}, \tag{A.13}$$

where

$$b_{\mathrm{sm}} = -\alpha^{pro}, \tag{A.14}$$

$$d_{\mathrm{sm}} = 2\alpha^{pro}v_{\mathrm{H}} + 2(1-\alpha^{pro})\Pr(u=1)(v_{\mathrm{H}} - v_{\mathrm{L}}), \tag{A.15}$$

$$e_{\mathrm{sm}} = -\alpha^{pro}(v_{\mathrm{H}})^2 - 2(1-\alpha^{pro})\Pr(u=1)(v_{\mathrm{H}} - v_{\mathrm{L}})\mathrm{E}[v], \tag{A.16}$$

and the following closed form formula for the pooling competitive TTC

$$TTC_{\mathrm{pc}} = \frac{-d_{\mathrm{pc}} + \sqrt{d_{\mathrm{pc}}^2 - 4b_{\mathrm{pc}}e_{\mathrm{pc}}}}{2c_{\mathrm{pc}}} - q, \tag{A.17}$$

where

$$b_{\mathrm{pc}} = -[\beta\alpha^{pro} + (1-\beta)\alpha^{non}], \tag{A.18}$$

$$d_{\mathrm{pc}} = 2[\beta\alpha^{pro} + (1-\beta)\alpha^{non}]v_{\mathrm{H}} + 2[\beta(1-\alpha^{pro}) + (1-\beta)(1-\alpha^{non})]\Pr(u=1)(v_{\mathrm{H}} - v_{\mathrm{L}}), \tag{A.19}$$

$$e_{\mathrm{pc}} = -[\beta\alpha^{pro} + (1-\beta)\alpha^{non}](v_{\mathrm{H}})^2 - 2[\beta(1-\alpha^{pro}) + (1-\beta)(1-\alpha^{non})]\Pr(u=1)(v_{\mathrm{H}} - v_{\mathrm{L}})\mathrm{E}[v]. \tag{A.20}$$

These analytic formulae are using in Sections 3.3 and 3.4.

*Proof of Proposition 5. Path 1*(a): The maximum value of $a_{\mathrm{pm}}$ is $v_{\mathrm{H}}$. Hence, the maximum value of $etc_{\mathrm{pm}}$ is $v_{\mathrm{H}} - q - \varepsilon$, which is less than $etc_{\mathrm{mn}} = v_{\mathrm{H}} - q$. (b) The minimum value of $ETC_{\mathrm{pm}}$ is $a_{\mathrm{sm}} - q = \mathrm{E}[v \mid t \in pro, TTC_{\mathrm{sm}}] - q$, which is greater than $TTC_{\mathrm{pc}} = \mathrm{E}[v \mid t \in both, TTC_{\mathrm{pc}}] - q$. The fact that $TTC_{\mathrm{sm}} > TTC_{\mathrm{pc}}$ follows from the fact that professional orders are going to have a higher conditional mean than a pooled average of professional and nonprofessional orders.

*Path 2*(a): We know that $TTC_{\text{pm}} = v_{\text{H}} - q$ and we know that $TTC_{\text{sm}} < v_{\text{H}} - q$, due to the fact that the expression for $TTC_{\text{sm}}$ has a positive weight on $\text{E}[v]$. $PI_{\text{sm}}^{\text{third-market}} < PI_{\text{sm}}^{\text{primary market}}$ follows from the fact that non-professional orders which go to the third market are less likely to be informed than professional orders which go to the primary dealer. (b) $TTC_{\text{sm}}$ is arbitrarily close to $a_{\text{sm}} - q$, which is greater than $TTC_{\text{pc}}$ (see Path 1(b) above). The probability of informed trading being lower in the third market follows from the fact that the nonprofessional orders which go to the third market are less likely to be informed than professional orders which go to the primary dealer.

*Path 3*(a): $TTC_{\text{sm}}$ is arbitrarily close to $a_{\text{sm}} - q$, which is greater than $TTC_{\text{pc}}$ (see Path 1(b) above). The probability of informed trading being lower in the third market follows the same reasoning as Path 2(a) proof above. (b) This is the same as the Path 2(b) proof above.

*Proof of Proposition 6.* The joint probabilities of each of the eight possible events are:

$$q_1 = \beta\gamma\alpha^{ps}, \quad q_2 = \beta\gamma(1-\alpha^{ps}), \tag{A.21}$$

$$q_3 = \beta(1-\gamma)\alpha^{pn}, \quad q_4 = \beta(1-\gamma)(1-\alpha^{pn}), \tag{A.22}$$

$$q_5 = (1-\beta)\delta\alpha^{ns}, \quad q_6 = (1-\beta)\delta(1-\alpha^{ns}), \tag{A.23}$$

$$q_7 = (1-\beta)(1-\gamma)\alpha^{nn}, \quad q_8 = (1-\beta)(1-\gamma)(1-\alpha^{nn}). \tag{A.24}$$

The expressions for the weights and probability are:

$$w_{cib}^{ps} \equiv \frac{q_1 \Pr[v \geqslant q + TTC_{cib}]}{q_1 \Pr[v \geqslant q + TTC_{cib}] + q_2[\Pr(u=1)]}, \tag{A.25}$$

$$w_{scib}^{nops} \equiv \frac{(q_3 + q_5 + q_7)\Pr[v \geqslant q + TTC_{vib}]}{(q_3 + q_5 + q_7)\Pr[v \geqslant q + TTC_{cib}] + (q_4 + q_6 + q_8)\Pr[u=1]}, \tag{A.26}$$

$$\Pr(t \notin ps \mid buy, TTC_{cib}) \equiv \frac{(q_3 + q_5 + q_7)\Pr[v \geqslant q + TTC_{vib}] + (q_4 + q_6 + q_8)\Pr[u=1]}{(q_3 + q_5 + q_7)\Pr[v \geqslant q + TTC_{cib}] + (q_4 + q_6 + q_8)\Pr[u=1] + q_1 \Pr[v \geqslant n_{ssc}] + q_2[\Pr(u=1)]}. \tag{A.27}$$

*Competitive internalization both* (*cib*) *case*: First, we consider a super-separating equilibrium in which brokers internalize all orders except professional sophisticated orders. In this case, the probability $f = \Pr(t \notin ps \mid buy,\ TTC_{cib})$ and is calculated as the ratio of the probability of the six events (p.n. inf. buy, p.n. uninf. buy, n.s. inf. buy, n.s. uninf. buy, n.n. inf. buy, and n.n. uninf. buy,) which do *not*

yield professional sophisticated buys divided by the probability of all eight events (adding p.s. inf. buy and p.s. uninf. buy) which yield a buy. The profit margin is the difference in conditional means $\pi = a_{cib} - m_{cib}$.

Substituting into $c = -f\pi$ yields the competitive commission. The competitive ask equals the conditional mean given professional orders $a_{cib} = \mathrm{E}[v \mid t \in ps, TTC_{cib}]$ and is calculated in the same spirit as Proposition 2. The conditional mean $m$ equals the conditional mean given nonprofessional orders $m_{cib} = \mathrm{E}[v \mid t \notin ps, TTC_{cib}]$ and is calculated in the same spirit as Proposition 2.

Next, consider the possibility of internalizing all types and offering a commission equal to the minus the difference in conditional means. This is not an equilibrium because individuals have an incentive to deviate from it. Specifically each individual broker has an incentive to earn positive rents by offering a slightly smaller commission and internalizing only orders that are not professional sophisticated.

*Monopolistic internalization both (mib) case*: The competitive internalizers have a second-mover advantage over the monopolistic primary dealers. Any attempt by the monopolistic primary dealers to earn positive rents will result in having that order flow internalized too. Hence, the monopolistic primary dealer is forced to price competitively or not participate. No participation can be ruled out based on the reasons discussed above. That is, pooling will not lead to an equilibrium because individuals have an incentive to deviate. Hence, the only equilibrium comes when the monopolistic primary dealer prices competitively and thus all the CIB results are reproduced in the MIB case.

*Total trading cost*: To show that $TTC_{ssc} = TTC_{comp}$, take the expression for $a_{cib}$, subtract $q$, and add the expression for $c_{cib}$, then combine and simplify the result and one obtains the expression for $TTC_{\mathrm{comp}}$ in Lemma 1. □

## References

Ahn, H., Cao, C., Choe, H., 1996. Tick size, spread, and volume. Journal of Financial Intermediation 5, 2–22.

Battalio, R., 1997. Third-market broker-dealers: cost competitors or cream skimmers? Journal of Finance 52, 341–352.

Battalio, R., Greene, J., Jennings, R., 1997. Do competing specialists and preferencing dealers affect market quality? Review of Financial Studies 10, 969–993.

Battalio, R., Hatch, B., Jennings, R., 2000. Dimensions of best execution for market orders: assessing differences between the NYSE and the Nasdaq Third Market. Working paper, Indiana University.

Bessembinder, H., Kaufman, H., 1997. A cross-exchange comparison of execution costs and information flow for NYSE-listed stocks. Journal Of Financial Economics 46, 293–319.

Bryan-Low, C., 2000. Web brokers begin to offer no-commission stock trades. Wall Street Journal.

Cheng, M., 1995. Payment for order flow and price improvement: The evolution of inter-market competition. Working Paper, University of California, Berkeley.

Chordia, T., Subrahmanyam, A., 1995. Market making, the tick size, and payment-for-order-flow: theory and evidence. Journal of Business 4, 543–575.

Cochrane, J., 1993. US equity market competitiveness. Working paper, New York Stock Exchange.
Easley, D., Keifer, N., O'Hara, M., 1996. Cream-skimming or profit-sharing? The curious role of purchased order flow. Journal of Finance 3, 811–833.
Glosten, L.R., 1991. Asymmetric information, the Third Market, and investor welfare. Working paper, Columbia University.
Glosten, L.R., Milgrom, P.R., 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.
Harris, L., 1997. Decimalization: A review of the arguments and evidence. Working paper, University of Southern California.
Huang, R., Stoll, H., 1996. Dealer versus auction markets: a paired comparison of execution costs on NASDAQ and the NYSE. Journal of Financial Economics 41, 313–357.
Kandel, E., Marx, L., 1999. Payments for order flow on NASDAQ. Journal of Finance 54, 35–66.
Lamoreux, C., Schnitzlein, C., 1997. When it's not the only game in town: the effect of bilateral search on the quality of a dealer market. Journal of Finance 52, 683–712.
Lee, C., 1993. Market integration and price execution for NYSE-listed securities. Journal of Finance 48, 1009–1038.
Lin, J., Sanger, G., Booth, G., 1995. Trade size and components of the bid-ask spread. Review of Financial Studies 8, 1153–1183.
Neal, R., Reiffen, D., 1994. The effect of integration between broker/dealers and specialists. Working paper, Indiana University Purdue University, Indianapolis.
NASD Payment for Order Flow Committee, 1991. Inducements for order flow. Report by the National Association of Securities Dealers, Washington, DC.
Porter, D., Weaver, D., 1997. Tick size and market quality. Financial Management 26, 5–26.
Ricker, J., 1997. Decimal pricing: nickel markets. Working paper, 1730 Filbert Street, No. 105, San Francisco, CA 94123.
U.S. Securities and Exchange Commission, 1997. Report on the Practice of Preferencing.