Does the Limit Order Routing Decision Matter?

Robert Battalio
University of Notre Dame

Jason Greene
Georgia State University

Brian Hatch
University of Cincinnati

Robert Jennings
Indiana University

We examine the impact deciding to route limit orders away from the New York Stock Exchange (NYSE) has on three dimensions of execution quality with methodologies controlling for market conditions and order submission strategies. Overall differences in limit order execution quality between regional stock exchanges and the NYSE are small, suggesting that the order routing decision may not affect retail limit order traders substantively. Conditioning on the distance between the limit order's price and prevailing quotes, however, reveals systematic differences in execution quality. This implies that brokers can strategically route limit orders to improve retail limit order execution quality.

Although limit orders typically comprise more than half of a broker's order flow, the literature examining the impact of the order routing decision on limit order execution quality is sparse.[1] This is in stark contrast to the extensive literature on market orders that is nearly unanimous in concluding that the New York Stock Exchange (NYSE) offers superior execution prices

Predecessor papers titled "Limit Order Executions on the Philadelphia Stock Exchange" and "A Comparison of Equity Limit Order Execution Quality Across Trading Venues" existed as early as April 1996. We thank Merrill Lynch and the Philadelphia Stock Exchange for data and an anonymous referee, Jeff Bacidore, Mike Ferguson, Larry Glosten (the editor), Craig Holden, Shane Johnson, Charles Jones, Tim Loughran, Maureen O'Hara (the executive editor), Mark Peterson, Jeff Smith, Steve Wyatt, workshop participants at the University of Cincinnati, Georgia State University, Norwegian School of Management, University of Notre Dame, U.S. Securities and Exchange Commission, and University of Southern California, and individuals at the 1999 American Finance Association meeting for comments on earlier versions of the article. Address correspondence to Robert Jennings, Finance Department, Kelley School of Business, Indiana University, Bloomington, IN 47405-1701, or e-mail: jennings@indiana.edu.

[1] To our knowledge, SEC (1997) is the only study of limit order execution quality across trading venues. They compare the execution quality of limit orders on the New York Stock Exchange with that on the regional stock exchanges (i.e., Boston Stock Exchange, Chicago Stock Exchange, Cincinnati Stock Exchange, Pacific Exchange, and Philadelphia Stock Exchange).

The Review of Financial Studies Spring 2002 Vol. 15, No. 1, pp. 159–194

[see Blume and Goldstein (1992), Lee (1993), Angel (1994b), Petersen and Fialkowski (1994), Ross, Shapiro, and Smith (1996), Bessembinder and Kaufman (1997), SEC (1997), Battalio, Greene, and Jennings (1998), Ready (1999), Bacidore, Battalio, and Jennings (2000), Battalio, Hatch, and Jennings (2000), and Bacidore, Ross, and Sofianos (2000)]. As a result, brokers making the "regular and rigorous" assessment of relative execution quality mandated by the Securities and Exchange Commission (SEC) and regulators examining market structure issues must either ignore limit orders or rely on analyses taking the across-exchange mix of limit orders as exogenous. We examine the impact of the broker's choice between regional stock exchanges and the NYSE on limit order execution quality in two settings that attempt to control for market conditions and order submission strategies.

We begin by examining the execution quality of Merrill Lynch retail limit orders around their decision to reroute order flow from two regional stock exchanges to the NYSE. Assuming that order submission strategies and market conditions do not change over time, this comparison should shed light on relative limit order execution quality and the efficacy of Merrill's routing decision. The limited scope of this natural experiment constrains our ability to draw general conclusions about differences in limit order execution quality. Consequently we develop a simulation methodology allowing a broader study of these issues. The simulation compares the execution quality of actual limit orders routed to regional stock exchanges with identical, hypothetical limit orders routed to the NYSE. By construction, this methodology controls for market conditions and order submission strategies and can be used to examine relative execution quality for *any* sample of on-the-quote and quote-improving limit orders routed away from the NYSE.[2] We evaluate three dimensions of limit order execution quality: (1) the frequency with which limit orders fill, (2) the speed with which executed limit orders fill, and (3) the economic performance of limit orders. Evidence of systematic differences in execution quality across trading venues would suggest that the limit order routing decision matters.

Limit order execution quality depends on the length of the limit order queue and the supply of market orders with which the limit orders interact. High-quality limit order executions are likely to result from short limit order queues exposed to numerous market orders. To the extent that different trading venues have different market order-to-limit order ratios, the lack of intermarket time priority (i.e., no central limit order book) creates the possibility that limit order execution quality differs across trading venues.[3] If the market

[2] On-the-quote limit orders seek to purchase (sell) shares at the best bid (offer) price. Quote-improving limit orders have limit prices between the contemporaneous best bid and offer prices.

[3] Time priority requires that the order submitted first at a given limit price should execute first. If time priority were imposed across markets, then the limit order routing decision would be irrelevant. Typically intramarket time priority is imposed on orders routed to a particular venue. Currently intermarket time priority does not exist in the U.S. equity market.

for order flow is frictionless and brokers route orders to maximize customers' execution quality, then execution quality would be equalized across venues without universal time priority. Macey and O'Hara (1997), however, suggest that frictions in the market for order flow may imply across-market differences in limit order execution quality. Practices such as internalization and payment for order flow, which divert orders from the NYSE, could diminish limit order execution quality in New York.[4] Dealers purchasing/internalizing market orders cannot, however, exclude others from interacting with the diverted orders. By mandating that retail limit orders have priority over professional interests on *individual* venues, the courts allow traders to interact with "captured" order flow.[5] Whether such frictions create opportunities to improve customers' limit order execution quality through strategic order routing decisions is an unresolved empirical issue that we examine in this study.

The extant research most closely related to our work is Harris and Hasbrouck (1996) and SEC (1997).[6] Both of these studies document limit order execution quality, but do not address whether orders sent to one venue might receive better executions elsewhere. Harris and Hasbrouck (1996) focus on comparing the economic performance of NYSE limit and market orders, but do not make intermarket comparisons. With 27% of the trades in NYSE-listed securities occurring off the exchange in 1999 [see NYSE (2000)], a complete analysis of limit order execution quality must include other markets. The SEC (1997) documents fill rates, execution waits, and measures how prices move after limit orders execute on the NYSE and on the regional exchanges. The SEC finds that limit orders fill more frequently on regional exchanges than they do on the NYSE, but that the NYSE's average execution wait is shorter than the regionals' wait for filled orders. Because the SEC (1997) computes intermarket execution-quality statistics taking order routing choices as given, one must be careful in making inferences from their results. For example, the NYSE's fill rate for on-the-quote limit orders (i.e., buy orders at the best bid price and sell orders at the best offer price) is

[4] Both practices allow brokers to share the dealer revenue associated with executing their customers' orders. With payment for order flow, brokers receive direct cash payments for sending order flow to the dealer. Brokers take the other side of customers' internalized orders. These practices are used to isolate uninformed *market* orders. Brokers, however, are reluctant to route market orders to one venue and limit orders to another for fear of attracting regulatory attention. Typically a routing destination is selected for orders in a stock conditional on order size. During the sample period some market makers paid for limit orders as well as market orders.

[5] See SEC Release no. 34-34753. This rule prohibits a market maker from trading for its own account at a price at which the market maker could execute a customer's limit order it is holding without executing the customer's limit order at the limit price (or better). This is known as the Manning rule (the name of the plaintiff bringing the initial suit against E. F. Hutton & Co.).

[6] Other limit order research focuses on traders' use of limit orders, the returns generated by limit order strategies, and limit orders' effects on markets. Kumar and Seppi (1993), Angel (1994a), Harris (1994), Chakravarty and Holden (1995), and Handa and Schwartz (1996) model a trader's decision to use limit orders. Handa and Schwartz (1996) and Harris and Hasbrouck (1996) find that limit order trading strategies may be profitable. Glosten (1994), Bias, Hillion, and Spatt (1995), Seppi (1997), Parlour (1998), and others examine how limit orders define and affect markets. Lo, MacKinlay, and Zhang (1997) model execution waits for limit orders using survival analysis.

45.5%, while 58.7–76.8% of these orders fill on regional exchanges. This suggests that on-the-quote limit orders are more likely to fill on the regionals than on the NYSE. The SEC does not ask, however, if regional orders might have enjoyed an even higher fill rate if they had been routed to New York. Brokers might send orders to regional exchanges with their customers' best interests in mind. If so, then identical NYSE orders would have provided execution quality no greater than that of the regional exchanges. It is possible, however, that conflicts of interest prevent customers from obtaining the better executions available in New York and/or that brokers route orders to the NYSE when market conditions suggest that it is more difficult to fill the order. Our experiments are designed to distinguish between these possibilities.

Our results suggest that on-the-quote limit orders receive better executions when routed away from the NYSE. Regional exchanges tend to fill on-the-quote limit orders more frequently and faster than does the NYSE. We also find that the opportunity cost of the foregone executions for on-the-quote limit orders routed to the NYSE is positive. Conversely we find that the NYSE provides superior execution quality to quote-improving limit orders. The NYSE offers faster, more frequent, and from the investor's perspective, more profitable executions than the regional exchanges for these orders. Together these results suggest that brokers can affect the quality of limit order executions with strategic routing decisions if they are willing to condition their order routing decision on the relationship of the limit price to the contemporaneous quoted price and send orders for the same stock to multiple market centers.

We describe our data in the next section. The event study is discussed in Section 2. In Section 3, we introduce our simulation methodology and document its accuracy. The simulations compare actual limit order outcomes with simultaneously submitted, identical *hypothetical* NYSE limit orders.[7] For the limit orders studied, the simulation correctly classifies more than 97% of the sample NYSE orders as filled or unfilled and accurately estimates execution times on average. To compute the value associated with differences in fill rates between two venues, we develop a measure of economic performance in Section 4. Sections 5, 6, and 7 examine the execution quality of limit orders on the Philadelphia Stock Exchange, Cincinnati Stock Exchange, and Boston and Pacific Exchanges, respectively, relative to our simulation methodology's estimate of the execution quality these orders would have obtained on the NYSE. Section 8 summarizes and concludes.

[7] More specifically, we compare the actual execution statistics to an exchange displaying the same quoted prices and sizes and receiving the same order flow as the NYSE but observing strict time priority with no hidden trading interests. The NYSE allows time priority deviations and the "crowd" may choose not to display trading interests. We demonstrate in Section 3 that our approach accurately matches a sample of actual NYSE limit order executions.

1. Data and Descriptive Statistics

We wish to evaluate the importance of the limit order routing decision. To do so we compare limit order execution quality on four regional exchanges (Boston, Cincinnati, Pacific, and Philadelphia) to that on the NYSE. We employ two approaches in making this comparison. First, we exploit a natural experiment in which Merrill Lynch ceased routinely routing order flow to regional exchanges in favor of New York. Merrill Lynch provides order audit-trail data from the Boston, Pacific, and New York exchanges (the "Merrill sample"). We use these data to compare limit order executions before and after orders are rerouted. Because this requires comparisons across time, it is possible that this analysis may not adequately control for differences in order submission strategies and market conditions.

Our second approach attempts to hold constant market conditions and order submission strategies by comparing the execution quality of actual regionally routed limit orders to that of identical, simulated NYSE orders. For this examination, we obtain order audit trail data from the Philadelphia and Cincinnati Stock Exchanges in addition to further exploiting the Boston and Pacific exchange data from the Merrill sample. In our simulation we "submit" an identical, hypothetical NYSE limit order at the same time each regional exchange limit order is submitted and compare the execution quality of these paired orders.

1.1 Data description

The data include security symbol, order type, order submission date and time, order size and price, an indication of whether the order is buyer or seller initiated, a settlement code, and the order's time in force. The data also note if the order is canceled. For executed orders we know the date and time of the execution as well as the trade size and price. We examine nonmarketable limit orders (i.e., buy orders with limit prices less than the contemporaneous national best offer price and sell orders with limit prices exceeding the national best bid price). We exclude orders associated with trades reported late or out of sequence and orders for nonregular settlement. We include canceled orders.

Merrill Lynch data are from October and November 1995. Merrill's NYSE system orders are available electronically, but we hand gather its regional data from paper reports. This limits our regional exchange analysis of Merrill orders to October 2–6 and November 17, 27–30, 1995. In October, Merrill routed orders to the Boston, Pacific, and New York Stock Exchanges. All of its orders went to the NYSE in November. Philadelphia Stock Exchange data cover November 1, 1995–January 31, 1996. We obtain data from five Cincinnati Stock Exchange broker-dealers for January 20–24, 1997. Although our data sources produce disparate sample periods, our analyses individually examine each dataset. The event study compares the Merrill data in October to the Merrill data in November. The only known difference in the data between these periods is that orders routinely go to regional exchanges in October but to New York in November. Our simulations, likewise, compare the outcomes of actual regional limit orders from a particular exchange to identical, hypothetical NYSE limit orders within the sample period unique to those data. We cannot determine whether differences in relative (to the NYSE) execution quality between these samples are due to differences between regional exchanges or a time-series trend in relative execution quality. Our focus, however, is comparing the regional exchanges' execution quality to that of the NYSE, not comparing the regional exchanges to each other. We believe that our experimental design minimizes the effect of different sample periods.

1.2 Descriptive statistics

Table 1 summarizes our data sources and provides descriptive statistics of its properties. The Cincinnati and Merrill data include all limit orders during the sample periods. Philadelphia provides only nonmarketable orders. From the orders provided, we exclude orders for fewer than 100 shares (odd lots) because they are not recorded in the publicly available data used in the simulation. We also eliminate good-until-canceled orders, assuming that the trading strategies producing these orders differ from the strategies producing day orders. Day orders are 47% of the orders in Philadelphia (recall we did not obtain marketable limit orders), 62% of the Cincinnati orders, and 66% of Merrill's limit orders. Our analyses require valid quotes when the order is submitted, which we obtain from the TAQ files. We construct a national best bid and offer (NBBO) quote stream using the last recorded quote prior to the order's receipt as the benchmark quote. A valid quote has nonzero bid and offer prices and sizes, a national best bid (NBB) price less than the national best offer (NBO) price, and an NBBO spread less than $5.00. This eliminates orders received when the market is closed or when trading is suspended. Merrill's regional exchange data have only to-the-minute time stamps. We eliminate Merrill orders arriving during a minute in which either the NBB/O or NYSE bid/offer prices change because we classify the limit order's aggressiveness by comparing the limit price to the benchmark quoted price. This additional screen for the Merrill data increases the number of Merrill orders eliminated due to invalid quotes to more than 30% of the sample compared to 4.5% (7.9%) for the Philadelphia (Cincinnati) sample. To make the other samples comparable with the data from the PHLX, we eliminate marketable orders.[8] These screens leave 43,033 Philadelphia orders, 11,798 Cincinnati orders, and 3,467 Merrill orders.

[8] This restriction does not impose an artificial constraint on the final sample. We would eliminate marketable limit orders because they are hybrid orders; neither a market order nor a traditional limit order. We use an event study and simulation to investigate the quality of limit order executions across trading venues.

Table 1
Descriptive statistics

Panel A: Data screens

	Audit trail data		Merrill Lynch orders			
	PHLX (11/95–1/96)	CSE (1/96)	BSE (10/95)	PCX (10/95)	NYSE (10/95)	NYSE (11/95)
Initial sample of limit orders	101,831	34,093	1,541	1,864	4,416	9,951
GTC orders	54,183	12,983	504	710	1,214	3,694
Invalid quotes	4,615	2,693	356	314	1,790	2,976
Marketable orders	n.a.	6,619	303	371	568	1,505
Remaining day orders	43,033	11,798	378	469	844	1,776

Panel B: Descriptive statistics

	PHLX (11/95–1/96)	CSE (1/96)	BSE (10/95)	PCX (10/95)	NYSE (10/95)	NYSE (11/95)
Pre-noon orders (Eastern time)	50.80%	49.47%	41.58%	57.36%	44.55%	50.28%
Buy orders	55.36%	49.42%	44.71%	36.46%	49.41%	52.03%
Canceled orders	4.60%	15.51%	18.78%	17.70%	15.17%	21.40%
Fill rate	66.12%	50.19%	62.96%	63.33%	55.92%	56.12%
Order size (shares)	650	774	490	549	2,576	1,628
Depth (shares)	19,010	20,054	8,608	19,635	10,163	9,783
Limit price	$31.94	$44.07	$36.85	$35.69	$36.88	$36.80
Spread width	$0.1674	$0.1457	$0.1839	$0.1711	$0.1762	$0.1819
Execution time (minutes)	23.98	16.23	21.43	47.44	19.59	18.78

Panel C: Limit price relative to the relevant quote (number of orders, fill rate)

Distance from relevant quote	PHLX (11/95–1/96)	CSE (1/96)	BSE (10/95)	PCX (10/95)	NYSE (10/95)	NYSE (11/95)
Quote-improving	8,275	680	107	106	152	403
	89.21%	89.71%	81.31%	90.57%	93.42%	93.05%
On-the-quote	21,592	4,956	169	249	400	818
	78.10%	72.92%	62.13%	68.67%	64.25%	58.68%
1/8 away	5,355	1,974	43	45	146	220
	46.01%	42.40%	44.19%	40.00%	36.30%	32.73%
1/4 away	2,564	1,072	24	29	50	114
	34.05%	33.21%	58.33%	27.59%	26.00%	34.21%
3/8 away	1,521	681	10	9	29	55
	25.62%	29.07%	60.00%	22.22%	3.45%	25.45%

Table 1
(continued)

Panel C (continued)

Distance from relevant quote	PHLX (11/95–1/96)	CSE (1/96)	BSE (10/95)	PCX (10/95)	NYSE (10/95)	NYSE (11/95)
1/2 away	950	484	2	5	14	35
	22.74%	20.25%	50.00%	40.00%	14.29%	17.14%
More than 1/2 away	2,776	1,951	23	26	53	131
	9.80%	10.66%	26.09%	0.00%	7.55%	9.16%

PHLX represents the Philadelphia Stock Exchange. CSE represents the Cincinnati Stock Exchange. The Merrill Lynch orders execute on the Boston Stock Exchange (BSE), the New York Stock Exchange (NYSE), and the Pacific Exchange (PCX) in October and the NYSE in November. The PHLX data are collected from November 1, 1995 through January 31, 1996. The CSE data are collected from January 20 through January 24, 1996. The October Merrill Lynch data are collected from October 2 through October 6, 1995. The November Merrill Lynch Data are collected on November 17, 1995 and from November 27 through November 30, 1995. The data from the PHLX exclude marketable (i.e., buy orders with limit prices greater than or equal to the order-submission time national best offer price and sell orders with limit prices less than or equal to the order submission time national best bid price) limit orders. The other data sources contain all limit orders. We eliminate orders for fewer than 100 shares. GTC represents good until canceled. At the time each order arrives at the trading venue of interest we require that national best bid and offer (NBBO) prices exist (e.g., the market must be open), that the NBBO quoted prices have a positive size associated with them, and that the national best offer (NBO) price exceed the national best bid (NBB) price. In addition, because the Merrill Lynch data from the regional exchanges have only a to-the-minute time stamp, we require that both NBBO and NYSE quoted prices be stable during the minute of order arrival for these data. Marketable buy (sell) orders have limit prices greater (less) than or equal to the NBO (NBB) price. Because the PHLX data do not contain marketable limit orders, we eliminate them from all of the samples. The abbreviation "n.a." stands for not applicable. For buy (sell) orders, the relevant quote is the NBB (NBO) price at the time of order submission. Depth is the NBBO quoted size at the time of order submission in shares: bid (offer) for buy (sell) orders. Spread Width is the difference between the NBO and the NBB prices at the time the order arrives. The following seven lines sum to the number of sample orders: quote improving, on-the-quote, 1/8 away, 1/4 away, 3/8 away, 1/2 away, and greater than 1/2 away. Quote improving orders have limit prices between the national best bid and Offer price. On-the-quote buy (sell) limit orders have limit prices equal to the NBB (NBO) price. Other buy (sell) limit orders have limit prices less (greater) than the NBB (NBO) price by the indicated amount. The first number in each cell in panel C is the number of orders falling in that category. The second number in each cell in panel C is the fill rate for that category.

To conduct the descriptive statistical tests, we aggregate the October Merrill sample data across trading venues. This makes the October Merrill sample comparable to Merrill's November data. We cannot reject the null hypothesis that half of each sample's day limit orders arrive before noon Eastern time using a chi-squared test at traditional significance levels. About half of each sample's orders are buy orders except for the Merrill October sample, where significantly (at the .01 level) less than 50% of the orders are buys. Cancellation rates range from 4.6% in Philadelphia to 21.4% in Merrill's November data, both of which are significantly different from the October Merrill and Cincinnati samples. Merrill orders (both November and aggregated October) are of larger size and arrive when there is less NBBO quoted depth than Philadelphia or Cincinnati orders. Cincinnati orders have higher limit prices and narrower order submission time NBBO quoted spreads, which is probably because the Cincinnati Stock Exchange trades only widely followed stocks.

Requiring that orders execute completely to be considered filled produces fill rates ranging from 50% for Cincinnati to 66% in Philadelphia. The Cincinnati rate is significantly less than the rate from each other sample, and the Philadelphia figure is significantly greater than each other sample (again, aggregating Merrill's October data). Execution waits are significantly longer in the aggregated October Merrill data (due to the Pacific data) than in any other sample. Panel B also provides insight into the potential problems with making intermarket comparisons of execution quality. Except for Cincinnati, which we return to below, each regional exchange exhibits significantly higher fill rates than the NYSE (Boston at the .05 level, Philadelphia and Pacific at the .01 level). The average NYSE order size, however, is significantly larger than the order sizes on the regional exchanges. We require that all shares in an order execute to categorize the order as filled, which may explain fill rate differences.

Panel C reports the distribution of limit prices relative to the NBBO and the associated fill rates. As in Harris and Hasbrouck (1996), the most popular nonmarketable limit orders are on-the-quote orders (orders to buy at the NBB or sell at the NBO). Although the NYSE's fill rate for on-the-quote (hereafter, OTQ) orders exceeds the comparable fill rate in SEC (1997), the regional exchanges' fill rates exceed the NYSE's for these orders as in SEC (1997). In all but the Cincinnati data, the second most frequent order is a quote-improving order (one with a limit price between the current quoted prices). Contrary to SEC (1997), we find that the NYSE has the highest fill rate for quote-improving (hereafter, QI) orders. This difference in findings is probably because our samples contain only the most actively traded NYSE-listed stocks. Finally, we note that Cincinnati has competitive fill rates in most of the limit price categories. Why then is the CSE's overall fill rate (panel B) the lowest? Panel C also reveals that the Cincinnati limit order prices tend to be further from the NBBO than orders on the other market centers. For Merrill and Philadelphia orders, 69% of the limit orders are QI or OTQ orders. Those orders comprise only 48% of Cincinnati limit orders. This reinforces the idea that care must be taken when making unconditional, across-exchange comparisons of limit order fill rates.

2. A Natural Experiment

On October 20, 1995, the *Wall Street Journal* reported that Merrill Lynch would cease routinely sending small retail orders in stocks listed on the New York and American Stock Exchanges to Merrill-affiliated regional specialists.[9] At the time of the announcement, Merrill routed orders in stocks for which it acted as a Pacific Exchange (PCX) specialist to the PCX and orders in other stocks to the Boston Stock Exchange (BSE). These arrangements allowed Merrill to obtain dealer revenue associated with executing its market orders. By providing liquidity on the Pacific Exchange, Merrill specialists earned the effective half-spread for Merrill in addition to commissions. In exchange for sending orders to Boston, Merrill received order-flow payment.

[9] See *Wall Street Journal* (1995) and Battalio, Greene, and Jennings (1998). Conversations with Merrill officials verify the accuracy of this story. For the affected securities, the decision whether to route the order to the regional or the national exchange was based primarily on the order's size.

After the close of trading on October 31, 1995, all orders (market and limit) in the affected stocks went to New York, unless a regional posted a better quote. This move involved Merrill selling its Pacific specialist operations and terminating its Boston relationships. Before the rerouting, our sample data indicate that the BSE received 20% and the PCX received 24% of Merrill's limit orders (in different stocks), leaving about 56% for the NYSE. The rerouting is complete. No sample November limit orders go to the regional exchanges.

We use Merrill's policy change to conduct an event study with the 3,467 nonmarketable day limit orders in the affected stocks. We divide these orders into two samples. The "Boston sample" contains all orders in the stocks for which Merrill chose between Boston and NYSE routing in October regardless of the market center receiving the order. The "Pacific sample" includes all orders in the stocks for which it decided between the PCX and the NYSE in October regardless of the venue. Because Merrill routed each stock to one regional, these samples are mutually exclusive. The Boston sample contains 1,528 orders; 378 orders routed to Boston in October (see Table 1), 268 NYSE October orders, and 882 NYSE orders in November. The Pacific sample has 1,939 orders; 469 PCX October orders (see Table 1), 576 to New York in October, and 894 to New York in November. We compare fill rates and execution waits in October and November. To do so, we combine the regionally routed and the NYSE-routed orders for the October calculations in order to represent the entire package of limit orders in the affected stocks during that period. Hence the reported October figures represent a weighted average of regional and NYSE limit orders. All November orders receive NYSE routing. Because Merrill's routing choice considered order size during October, Table 2 reports the results of the analysis with order size categories determined by the automatic execution maxima on the relevant regional exchange.

Panel A reports execution-quality statistics for the Boston sample stocks. The fill rate for small QI orders routed exclusively to the NYSE in November is significantly greater at the .05 level than the analogous fill rate for orders routed to both the NYSE and Boston in October. Conversely, the fill rate for small orders with limit prices away from the NBBO falls significantly (again at the .05 level) after the switch to exclusive NYSE routing. Small OTQ orders experience almost an identical fill rate in the two periods. The overall small-order fill rate (not reported in Table 2) falls from 57% in October to 54% in November, a decrease that is not statistically significant at traditional levels. For executed small orders, execution waits are insignificantly shorter after the switch. Wait times generally are longer for large orders after the rerouting, although only the wait for orders with a limit price away from the NBBO is significantly longer. Our analysis shows that, on average, Boston sample stocks' October execution quality measures are comparable to the November execution quality measures. Because the major (known) difference

Table 2
The effect of Merrill Lynch's decision to reroute order flow away from regional exchanges to the New York Stock Exchange on limit order execution quality

Panel A: Orders in the stocks that Merrill Lynch routed *some* orders to the Boston Stock Exchange prior to November 1995

		Orders routed to both the BSE and the NYSE (October 1995)			Orders routed only to the NYSE (November 1995)		
		QI	OTQ	WTQ	QI	OTQ	WTQ
Orders for 100 to 1299 shares	Fill rate	84.67%	65.45%	39.20%	92.68%**	64.45%	29.30%**
	No. of orders	137	220	148	164	301	215
	Wait (minutes)	11.24	31.27	94.23	7.06	31.15	78.42
Orders in excess of 1299 shares	Fill rate	100.00%	42.11%	19.72%	84.62%	45.21%	15.53%
	No. of orders	13	57	71	26	73	103
	Wait (minutes)	9.98	52.05	22.89	18.77	51.80	103.53***

Panel B: Orders in the stocks that Merrill Lynch routed *some* orders to the Pacific Stock Exchange prior to November 1995

		Orders routed to both the BSE and the NYSE (October 1995)			Orders routed only to the NYSE (November 1995)		
		QI	OTQ	WTQ	QI	OTQ	WTQ
Orders for 100 to 1099 shares	Fill rate	91.53%	69.27%	25.00%	95.79%*	57.38%***	25.79%
	No. of orders	189	397	196	190	359	159
	Wait (minutes)	12.98	62.16	120.52	4.86*	37.24***	91.29
Orders in excess of 1099 shares	Fill rate	88.46%	62.50%	30.11%	82.61%	55.29%	29.49%
	No. of orders	26	144	93	23	85	78
	Wait (minutes)	40.84	57.72	90.44	8.51	51.78	113.92

The October Merrill Lynch data are collected from October 2 through October 6, 1995. The November Merrill Lynch Data are collected on November 17, 1995, and from November 27 through November 30, 1995. BSE represents the Boston Stock Exchange, NYSE is the New York Stock Exchange, and PCX is the Pacific Exchange. The fill rate is the number of completely filled orders divided by the total number of orders. A chi-squared test with one degree of freedom is used to conduct a test of difference in fill rates. One asterisk in a fill rate row denotes significance at the .10 level, two denote significance at the .05 level, and three denote significance at the .01 level. The wait time is the time elapsing between order submission and the complete filling of the order measured in minutes. It is computed only for completely filled orders. One asterisk indicates that the November mean wait time is statistically different at the .01 level from the analogous October rate using a two-sided t-test. Two (three) asterisks in a wait row indicate that the difference is significant at the .05 (.01) level. QI = Quote Improving. OTQ = On the Quote. WTQ = Worse Than the Quote.

between the two sample periods is the market center to which the majority of the small orders are routed, we conclude that the routing change has little impact on the average Merrill limit order trader in these stocks. We do find, however, that orders with limit prices inside of the existing NYSE quote fill more frequently in November, when they are routed exclusively to the NYSE, than in October, when the routing destinations include regional exchanges.

Execution quality for the Pacific sample stocks appears in panel B. The small QI (OTQ) orders' fill rate increases (decreases) significantly in November, suggesting that October's execution quality is lower (higher) than November's quality. These changes offset so the overall small-order fill rate (not reported) does not differ significantly at traditional levels between October and November. Execution times for small QI and OTQ orders fall (at the .10 significance level) between October and November, suggesting that the rerouting results in faster executions for filled orders. As with the Boston sample, small quote-improving orders in the Pacific sample stocks appear to execute more frequently and more quickly when routed to the NYSE than when they are routed to the PCX. On-the-quote orders, however, fill less frequently when routed exclusively to New York in November than in October when the routing decision included regional exchanges.

Assuming that (1) the order submission strategies of Merrill's traders using day limit orders are identical in October and November, and (2) market conditions are unchanged between the two sample months, the event study suggests that quote-improving orders fill more often and faster on the NYSE, but that regional exchanges offer execution quality at least as high as the NYSE for other limit orders. In fact, Boston limit orders with limit prices away from the contemporaneous quoted prices and Pacific limit orders with prices equal to the contemporaneous quoted price fill more frequently on the regional than on the NYSE. Although these results are provocative, the limited scope of the Merrill sample limits our ability to draw general conclusions about across-exchange differences in limit order execution quality. In addition, events like these do not occur often enough to track relative execution quality as mandated in SEC (1997). Consequently we develop a simulation methodology that estimates if/when limit orders fill while controlling market conditions and order submission strategies, show that the technique is accurate, and use the methodology to evaluate decisions to route limit orders away from the NYSE.

3. Simulation Methodology

In this section we develop a simulation methodology to estimate if/when certain types of limit orders execute on the NYSE.[10] Using actual NYSE limit orders as a benchmark, we test our process by submitting identical, hypothetical orders and comparing the actual orders' fill rates and execution waits to the hypothetical orders' statistics. The simulation is designed specifically for quote-improving and on-the-quote limit orders. Our calibration results demonstrate that the methodology accurately replicates the outcomes of our sample of actual orders.[11]

3.1 The simulation methodology

Our simulation presumes that the NYSE fully displays all public trading interests, enforces time priority equally for all trading interests, and follows

[10] Lo, MacKinlay, and Zhang (1997) use TAQ-recorded trades to derive upper and lower bounds on fill times for all limit orders, but find these bounds are too wide to be useful in forecasting actual execution times. We study only orders for which we can obtain additional information to more accurately estimate execution times. Data items we use to estimate fill times are included in their multivariate regressions.

[11] When testing our methodology, both the hypothetical and actual limit order exist on the NYSE. When using our methodology to examine relative execution quality, the actual order is on the regional while the hypothetical order resides on the NYSE. Although the Intermarket Trading System displays trading interests across exchanges, this difference may matter if the location on which trading interest is displayed affects traders' behavior.

strict time priority.[12] To determine the hypothetical order's standing in the limit order queue, we note the TAQ-recorded quote existing when we "submit" the hypothetical order. Because we need quoted size to estimate the submission time queue's length, we can examine only OTQ and QI orders. For each actual buy (sell) order we place an identical hypothetical order in line behind the existing orders on the bid side (offer side) of the submission time NYSE quote. For QI orders, price priority suggests that our order is first in the queue. For OTQ orders, time priority implies that our order is behind other orders already in line. After placing our order, we track subsequent TAQ-recorded quotes and trades. We presume our order executes when (1) the cumulative number of shares trading at the limit price after the order arrives equals or exceeds the submission time queue length plus the order's size,[13] (2) we observe a trade price less (greater) than the price at which we are willing to buy (sell), or (3) we find a revision in the submission time quoted price such that the bid (offer) price is less (greater) than our limit price and the benchmark quote. In the first case, enough shares trade to subsume our submitted order. The fill time associated with a hypothetical order presumed to execute in this fashion is the time of the trade in which the cumulative volume at the limit price since submission equals or exceeds the required volume. In the latter two cases, trading does not eliminate the shares in the limit order queue with higher time priority than our order. Apparently, traders with higher time priority cancel orders. Here the hypothetical order's fill time is the time of the last trade at the limit price prior to observing the inferior price. Orders (actual or hypothetical) must fill completely to be classified as filled.

An example may help illustrate our methodology. Suppose an actual 200-share buy limit order with a price of \$20 per share is submitted at 10:00 on a given day. The contemporaneous NYSE quote is \$20 bid for 4,000 shares and \$20.25 offered. Our data tell us if and when this actual order fills. We "submit" a hypothetical order to buy 200 shares at \$20 each at 10:00, assume that the simulated order queues behind the 4,000 shares on the limit order book

---

[12] McInish and Wood (1995) question the first assumption by finding evidence consistent with the claim that specialists do not display all public limit orders. Since then the NYSE has modified its rules so that specialists have less discretion in displaying public limit orders (NYSE Rule 79A) and has disciplined firms not fully displaying public orders [e.g., see Dow Jones News Service (2000)]. New York Stock Exchange (2000), Sofianos and Werner (1997), and Kavajecz (1999) note the importance of specialists and floor brokers in the NYSE's quoting and trading processes. Finally, NYSE Rule 72 allows violations of strict time priority. Violations of the first and third assumptions suggest that our imputed queue length is too short and violations of the second assumption suggest that our estimated queue length is too long. Partitioning the sample by attributes extant research finds related to specialist participation in the quoting/trading process does not substantively change the conclusions from those reported in the article. Modifying the simulation to allow for large orders to "size-out" the limit order book also does not change our conclusions.

[13] We also note apparent cancellations of orders with higher time priority than ours. For example, if we submit an OTQ order and find that the subsequent relevant quoted size is less than our estimated queue length, then we presume that orders with higher time priority than ours canceled and move our order that much closer to the front of the queue. Based on Hasbrouck, Sofianos, and Sosebee (1993), we subtract six seconds from the TAQ-recorded trade times.

at $20, and begin tracking quotes and trades. The order executes when the 4,200th (4,000 shares in the bid plus the 200-share order size) share trades at $20 on the NYSE after 10:00. So if we observe a 2,000-share trade at $20 at 10:12, another 2,000-share trade at $20 at 10:14, and a $20, 200-share trade at 10:18, then we assume the hypothetical order fills at 10:18. Alternatively, we may find a trade or bid price less than $20. If an inferior price is observed, then we assume that the hypothetical order executes at the time of the last trade at $20 prior to the inferior price (or when the order is submitted if no trade at $20 is observed between when the order is submitted and when we observe the inferior price). Again, suppose a 200-share limit buy order arrives at 10:00 with a $20 limit price. If we observe a trade at $20 at 10:10 and the bid price falls to (or a trade executes at) $19.875 at 10:12, then we assume our order executes at 10:10. Our hypothetical order's outcome (an execution at 10:18 or 10:10 in the examples above) is compared to the actual order's outcome recorded in our sample data.[14] We calculate differences in fill rates, fill waits, and economic performance measures to determine the ability of the simulation methodology to mimic actual outcomes.

3.2 Tests of the simulation methodology's accuracy

To test our methodology, we simulate the outcomes of a sample of actual NYSE orders as if they are hypothetical orders. We use OTQ and QI day limit orders that Merrill's customers submit throughout November 1995.[15] We know actual submission times and (if executed or canceled) execution or cancellation times for these 12,581 orders (5,237 QI and 7,344 OTQ). For each order we determine if/when our methodology indicates the order fills and compare the simulated outcome with the known actual outcome.

Descriptive statistics comparing the actual and hypothetical NYSE orders' fill rates are displayed in Table 3 by order size and type. The fill rate for hypothetical QI orders exceeds the actual rate by about one percentage point. For OTQ orders, the hypothetical fill rate is within 0.62% of the actual rate. Chi-squared tests of differences in proportions cannot reject the null hypothesis that the hypothetical fill rate equals the actual rate for any order-type/size category. In addition, the simulation correctly classifies 97.56% of the hypothetical orders as filled or unfilled. Finally, it is almost equally likely for the simulation to incorrectly classify an OTQ order as filled or not (particularly for the smaller order sizes). For QI orders, the simulation is more likely to execute orders not actually filling than it is to not fill an executed actual order, but it correctly classifies 98.64% of the orders as filled or not.

We report the distribution of differences in execution wait times for the 8,428 order pairs (the hypothetical order paired with the identical actual

[14] If the buy order's price is $20.125, then our order is at the front of the queue and fills when 200 shares trade at $20.125 (or when we observe a transaction price less than $20.125 or a bid price less than $20.00).

[15] We are not constrained by the limitations of the regional exchange data, so we examine all qualifying NYSE limit orders during the entire month of November.

Table 3
Comparisons of fill rates between actual and paired hypothetical New York Stock Exchange limit orders

Limit order size (shares)	Limit order type	Number of orders	Hypothetical execution type			Actual fill rate (%)	Hypothetical fill rate (%)	Accuracy statistics		
			Shares	Trade	Quote			Agree	False negative (%)	False positive (%)
All	QI	5,237	95.86	3.93	0.21	82.20	83.33	98.65	0.11	1.24
All	OTQ	7,344	85.11	3.74	11.15	58.14	57.52	96.79	1.92	1.29
100–599	QI	3,111	96.09	3.65	0.26	86.95	87.95	98.75	0.13	1.12
100–599	OTQ	3,496	86.26	3.20	10.55	64.39	63.47	96.28	2.32	1.40
600–1,099	QI	1,076	95.76	4.24	0.00	79.74	81.04	98.51	0.09	1.39
600–1,099	OTQ	1,668	87.35	2.57	10.08	59.11	58.27	97.12	1.86	1.02
1,100–5,099	QI	885	95.70	3.99	0.31	72.20	73.56	98.42	0.11	1.47
1,100–5,099	OTQ	1,683	82.40	5.71	11.89	49.73	49.97	97.03	1.37	1.60
>5,099	QI	165	93.33	6.67	0.00	62.42	63.64	98.79	0.00	1.21
>5,099	OTQ	497	72.47	7.29	20.23	39.44	38.63	98.39	1.21	0.40

The actual limit orders are Merrill Lynch orders submitted to the New York Stock Exchange via SuperDOT during November 1995. QI represents quote improving and OTQ represents on the quote. A QI limit order has a limit price between the contemporaneous bid and offer prices. An OTQ limit order is a buy (sell) limit order with a limit price equal to the contemporaneous bid (offer) price. Hypothetical orders in the "shares" column execute because the number of shares executing at the limit price after the submission time equals the depth associated with the limit price at the time the limit order is submitted plus the size of the order. Hypothetical orders in the "trade" ("quote") column execute because we observe a transaction (revised quoted) price that is inferior to the limit price (order submission time quoted price). The fill rate is the number of orders completely filled on the day submitted divided by the number of orders submitted. Agree indicates that the simulation correctly classify the order as filled or unfilled. False positive indicates that the simulation concludes a fill when the actual order did not fill. False negative indicates that the simulation finds that the order did not fill when it did.

order) with both orders filling in Table 4. Regardless of order size/type, the median difference between estimated and actual fill times is only three or four seconds, with the hypothetical order executing before the actual. This difference is associated with a mean wait time of about 1.5 minutes for quote-improving orders (a 3.4% error). OTQ orders have a mean wait time of nearly 16 minutes, so that the three-second difference represents an error of 0.3%. The hypothetical order fills after the actual order (a negative wait time difference) in only 22% of the order pairs. For a few OTQ orders, the hypothetical order fills considerably after the actual order (the 10th percentile of the distribution is about 7.5 minutes). In more than 82% of all executed order pairs, the hypothetical order executes within one minute of the actual.

3.3 Summary

Overall, our simulation methodology produces an estimated NYSE fill rate within one percentage point of the actual NYSE fill rate and properly fills or does not fill the hypothetical order in 97.56% of the cases. Although the wait time differences between actual and hypothetical OTQ orders can be large, the overall median execution wait time differential is only three seconds. In Sections 5, 6, and 7, we compare actual regional exchange limit order trade quality to the estimated execution quality that these orders would have obtained if they had been sent to the NYSE. The calibration results above suggest that the outcomes of simulated NYSE orders are estimated accurately.

4. Estimating Economic Performance

Fill rates and execution waits are descriptive statistics associated with limit order trade quality. Typically we assume that filled orders are preferred to unfilled orders and that, if filled, shorter waits are better than longer waits. This ordering, however, might not take us as far as we wish when evaluating execution quality. We might, for example, wish to estimate the economic impact of an order routing decision on the trader. To do so we compute the aggregate value of the positions taken when routing limit orders to one venue and compare that to the value of the positions acquired when routing the same limit order package to an alternative venue. Suppose that we simultaneously send 100 identical limit orders to two market centers and that venue A fills all of the orders filled at venue B plus 10 orders that venue B does not fill. Clearly venue A's fill rate exceeds venue B's fill rate by 10 percentage points. We wish to estimate what the additional fills are worth to the trader. If these 10 trades are profitable to the trader, then the fill rate difference is economically meaningful. Alternatively, the 10 trades may mean little to the trader economically. Subtracting the total one-day dollar return on the security positions acquired by routing the orders to venue B from the total one-day dollar return on the positions acquired by routing the orders to

Table 4
Distribution of execution time differences between actual and paired hypothetical New York Stock Exchange limit order executions

Limit order size (shares)	Limit order type	Number of orders	Percentiles in the distribution of difference between actual execution time and hypothetical time (seconds)								
			10%	20%	30%	40%	50%	60%	70%	80%	90%
All	QI	4,299	0.0	2.0	3.0	3.0	3.0	4.0	4.0	4.0	14.0
All	OTQ	4,129	−451.2	−35.0	−1.0	3.0	3.0	4.0	4.0	9.0	42.0
100–599	QI	2,701	−1.0	2.0	3.0	3.0	3.0	4.0	4.0	4.0	13.0
100–599	OTQ	2,170	−470.4	−32.2	−1.0	3.0	3.0	4.0	4.0	10.0	49.0
600–1,099	QI	857	0.0	2.0	2.0	3.0	3.0	4.0	4.0	4.0	10.4
600–1,099	OTQ	955	−349.4	−32.2	0.0	3.0	3.0	4.0	4.0	6.0	31.0
1,100–5,099	QI	638	0.0	3.0	3.0	3.0	4.0	4.0	4.0	5.0	19.3
1,100–5,099	OTQ	814	−366.5	−41.8	−1.0	2.0	3.0	4.0	4.0	7.0	38.0
>5,099	QI	103	−1.6	2.0	3.0	3.0	3.0	4.0	4.0	8.2	30.6
>5,099	OTQ	190	−858.0	−184.8	−7.3	0.6	3.0	3.0	4.0	4.0	50.8

The actual limit orders are Merrill Lynch orders submitted to the New York Stock Exchange via SuperDOT during November 1995. QI represents quote improving and OTQ represents on the quote. The sample for this table are the order pairs in which both the actual order and the paired hypothetical order execute. A QI limit order has a limit price that is between the contemporaneous bid and offer prices. An OTQ limit order is a buy (sell) limit order with a limit price equal to the contemporaneous bid (offer) price. The difference in actual and hypothetical execution times equals the actual execution time minus the hypothetical execution time.

venue A measures the economic advantage of venue A over venue B. This advantage can be substantial when the stock's price briefly reaches the limit price and then moves away. If venue A consistently executes orders under these circumstances and venue B does not, then its fill rate advantage over venue B is economically meaningful.

Harris and Hasbrouck (1996; hereafter HH) provide a limit orders' economic performance metric. Our economic performance measure, however, differs from their *ex ante* measure. HH examine performance differences between two trading strategies; one using limit orders and the other using market orders. Implicit in the HH *ex ante* measure is the assumption that the stock is traded regardless of price. In this article, we examine the performance of identical limit orders routed to different market centers. Investors using limit orders may be more patient than the HH traders. Our economic performance measure is similar to the HH *ex post* measure. Specifically, we compute the implicit gain or loss from a filled limit order evaluated at the closing price of the day it executes. For a filled limit buy (sell) order, we assume that the investor sells (buys) at the closing quoted bid (offer) price on the day the order is submitted. This figure represents the one-day dollar return on a filled limit order.

In Sections 5, 6, and 7, we compare the end-of-day portfolio value for the actual regionally routed orders to the estimated end-of-day portfolio value for the hypothetical NYSE orders. If both the actual and hypothetical orders fill, then our economic performance measure for each order in the pair is the same. Likewise, opportunity losses for order pairs in which neither order fills are equal. Of interest are order pairs where one order fills and the other does not. The implicit one-day gain (loss) for the filled order in the pair measures the economic advantage (disadvantage) of the venue filling the order. We compute this metric for order pairs in which the actual order fills and the hypothetical order does not and order pairs for which the hypothetical order fills and the actual order does not. Subtracting the return associated with the order pairs with only the hypothetical order filling from the return associated with the order pairs with only the actual order filling provides an estimate of the economic (dis)advantage of regional exchange routing relative to the NYSE. Dividing the aggregate value difference by the *total* number of shares in all the submitted orders provides the per share economic value of the routing decision for the sample package of limit orders.

Before making inferences about the economic performance of the hypothetical NYSE orders, we should determine how accurately the simulation methodology estimates this metric. To do so, we take the same approach (with the same data) used in the previous section to evaluate the simulation's accuracy in estimating fill rates and execution waits. Table 5 reports our economic performance metric for the Merrill order pairs in which one order fills and one does not (i.e., either the hypothetical order fills and the actual does not or the actual order fills and the hypothetical does not). These figures

Table 5
Comparisons of economic performance of actual and paired hypothetical New York Stock Exchange limit order executions

		Actual fills, hypothetical does not fill			Hypothetical fills, actual does not fill				
Limit order size (shares)	Limit order type	Orders	Total return ($)	Mean return per share ($)	Orders	Total return ($)	Mean return per share ($)	Total return difference ($)	Per share differences ($)
All	QI	6	1,048.25	0.2565	65	12,231.75	0.1488	−11,183.50	−0.0019
All	OTQ	141	24,845.88	0.1128	95	10,219.13	0.0788	14,626.75	0.0010
100–599	QI	4	548.25	0.5048	35	2,153.13	0.2109	−1,604.88	−0.0017
100–599	OTQ	81	3,190.75	0.1260	49	0.63	0.00	3,190.13	0.0030
600–1,099	QI	1	500.00	0.5000	15	203.75	0.0140	296.25	0.0003
600–1,099	OTQ	31	950.00	0.1569	17	4,662.50	0.3093	−3,712.50	−0.0025
1,100–5,099	QI	1	0.00	0.00	13	−600.00	−0.0150	600.00	0.0003
1,099–5,099	OTQ	23	14,580.1	0.2311	27	5,556.00	0.0676	9,024.13	0.0019
>5,099	QI	0	n.a.	n.a.	2	10,475.00	0.6055	−10,475.00	−0.0058
>5,099	OTQ	6	3,125.00	0.0293	2	0.00	0.00	3,125.00	0.0004

The actual limit orders are Merrill Lynch orders submitted to the New York Stock Exchange via Super DOT during November 1995. QI represents quote improving and OTQ represents on the quote. We examine economic performance only for those order pairs in which exactly one order in the pair executes. A QI limit order has a limit price that is between the contemporaneous bid and offer prices. An OTQ limit order is a buy (sell) limit order with a limit price equal to the contemporaneous bid (offer) price. To compute return, we compare the end-of-day national best bid (offer) price with the transaction price of an executed buy (sell). Filled buy (sell) orders in which the price rises (falls) have positive returns. We aggregate over all shares contained in orders in the indicated category. Mean return per share is the total return from the adjacent column divided by the number of shares used to calculate total return. The total return difference is the return when the actual fills and the hypothetical does not less the return when the hypothetical fills and the actual does not. The per share difference is the total return difference divided by the total shares in the indicated order type/size category.

do not represent the value difference between alternative routing decisions because both the actual and hypothetical order go to the NYSE. Instead, these numbers allow us to assess bias in our simulation methodology.

All but one of the returns are positive and most are small.[16] To measure net economic performance, we subtract the total return associated with the order pairs filling hypothetically, but not actually from the total return associated with the pairs filling only the actual order. This is the net economic performance for the package of orders submitted, because order pairs with both orders filled or both orders unfilled have zero net economic value. For our test, this is a package of 5,237 QI orders and 7,344 OTQ orders (Table 3) with a net economic performance of \$3,443.25 (= \$1,048.25 − \$12,231.75 + \$24,845.88 − \$10,219.13). Thus the net economic performance of the 12,581 orders and their hypothetical counterparts is \$0.2737 per order. With an average order size of 1,628 shares (Table 1), this is only \$0.00017 per share, suggesting that the simulation methodology provides an accurate estimate of the economic impact of filled orders. We now turn to comparing regional exchange limit order execution quality to the execution quality we estimate these orders would have received if routed to the NYSE instead of a regional exchange.

5. Exchange Execution Systems—Philadelphia Stock Exchange Rule 229.10

The Philadelphia Stock Exchange Automated Communication and Execution (PACE) System provides automatic order routing and execution. Although all exchanges have automated order routing systems, PACE includes apparently unique provisions for limit orders.[17] PHLX Rule 229.10 mandates that, upon obtaining time priority on the PHLX, 100- to 599-share limit orders with prices OTQ or worse in stocks listed on the national exchanges are due executions no later than when 1,000 shares of the stock trade at the limit

[16] This is not surprising. A negative return results from executed orders where traders experience regret (e.g., buying before a price decline). It is unlikely that either the hypothetical or actual order fails to fill in these cases. Suppose we submit a buy order with a limit price equal to the best bid price of \$30 when the best offer is \$30.25. Further suppose that the security's ask price falls to \$28 after the order is submitted and the order is not canceled. Even if sufficient shares do not trade at \$30 to trigger a hypothetical fill, our simulation executes the hypothetical order when we observe a (revised) quoted or trade price less than \$30. It also seems unlikely that the actual order fails to fill. Analogously, if the bid price in the above example quickly rises to \$32, then neither the actual nor the hypothetical order is likely to fill. Situations in which the price changes dramatically after order submission are unlikely to be included in our economic performance evaluation because both orders fill or both remain unfilled. This suggests that the order pairs likely to be included in our performance measure are those associated with small positive returns. Thus our economic performance measure, by examining only order pairs with exactly one order filling, is likely to differ from the negative performance measures of Harris and Hasbrouck (1996) and SEC (1997) that include all filled orders.

[17] Agency orders of less than a predetermined maximum size in stocks with bid prices greater than or equal to \$1.00 qualify for PACE. Brokers can request their orders not be subject to these execution rules.

price or better on the primary exchange.[18] For example, suppose an NYSE-listed stock's NBB is \$20 and its NBO is \$20.25. Rule 229.10 applies to small Philadelphia buy (sell) limit orders with prices less (greater) than or equal to \$20 (\$20.25) having time priority on the PHLX. A buy order at the front of the Philadelphia execution queue with a limit price of \$20 (sell order with a price of \$20.25) is due an execution either when market sell (buy) orders of sufficient size arrive in Philadelphia or when 1,000 shares trade at that price in New York, whichever occurs first. Although executions are manual, PACE notifies PHLX specialists that an eligible limit order is due an execution each time New York trades for at least 1,000 shares accumulate at the limit price. This 1,000-share queue is independent of the actual submission time quoted size at that price in New York. Thus, Philadelphia customers may move ahead of traders placing limit orders in New York (although it is possible that the PHLX trader may do worse than NYSE traders if there are many orders at the limit price in Philadelphia and little quoted size in New York). According to Rule 229.10, an order with a limit price bettering the existing NBB(O), for example, an order with a limit price of \$20.125, becomes part of the NBBO and is due an execution when the first opposite-side trading interest of sufficient size arrives in Philadelphia or (on a share-for-share basis) when trades occur at that price in New York.

5.1 Fill rates, economic performance, and execution waits of limit orders: PHLX versus NYSE

In Table 1 there are 43,033 Philadelphia nonmarketable, day limit orders with valid reference quotes. To assess whether there is an advantage to placing these orders in Philadelphia rather than on the NYSE, we compare the results of actual Philadelphia orders to identical hypothetical NYSE orders. We use the TAQ-recorded NYSE quote when an actual PHLX day limit order arrives to estimate the NYSE queue length. Thus we are constrained to evaluating orders with limit prices that either equal or better the NYSE quote. The Philadelphia data contain 9,558 day limit orders that are quote-improving relative to the contemporaneous NYSE quoted price and 20,667 limit orders that are on-the-quote relative to the NYSE quote.[19] Almost 70% of these 30,225 orders (21,014) qualify for PHLX Rule 229.10. For each sample order we simultaneously place an identical hypothetical order at the back of our

[18] This PHLX rule is designed to encourage brokers to route retail orders to Philadelphia. If brokers do not route market and limit orders to different trading venues, then the rule may increase the PHLX's market share of both order types. To our knowledge, the 1000-share level is arbitrary.

[19] The number of PHLX QI and OTQ orders here differs from that in Table 1. In Table 1 we use the NBBO to classify orders, but here we use the NYSE quote because we submit the hypothetical order to the NYSE. The number of NYSE QI orders exceeds the number of NBBO QI orders because the NYSE is not always at the NBBO. Compared to Table 1's PHLX column, this results in 1,283 more QI orders and 925 fewer OTQ orders. That is, 1,283 Philadelphia orders with limit prices OTQ (or worse) relative to the national best quoted price are QI relative to the NYSE quoted price. Correspondingly, there are 925 fewer OTQ orders relative to the NYSE quotes than relative to the NBBO.

estimated NYSE book. Our hypothetical NYSE order executes based on the algorithm described in Section 3. This allows us to estimate the improvement (if any) in the likelihood of filling, the wait for execution, and the economic performance offered by Philadelphia relative to the NYSE.

Table 6 reports fill rates and waits and performance measures for the 30,225 sample order pairs (an order pair consists of an actual PHLX order and an identical hypothetical NYSE order). Panel A reports fill rates. Of the 6,618 OTQ orders unfilled on both the PHLX and the NYSE, 1,880 orders execute in Philadelphia but not (given our estimation technique) in New York, 561 orders execute in New York but not in Philadelphia, and 4,177 execute on neither venue.[20] This implies that the PHLX's fill rate for the 20,667 OTQ orders is 77.1% ($= [20{,}667 - 561 - 4{,}177]/20{,}667$) and New York's rate is 70.7%. For the 13,988 OTQ orders subject to Rule 229.10, Philadelphia fills 85% compared to 74% in New York. The NYSE fill rate exceeds the PHLX's rate for large OTQ orders. Each difference is statistically significant at the .01 level. Finding that Philadelphia outperforms the NYSE for eligible orders but not for large orders is consistent with the claim that PHLX Rule 229.10 offers qualifying OTQ limit order traders opportunities to improve upon fill rates they might receive on the NYSE. The NYSE fill rate for QI orders significantly exceeds Philadelphia's for both order size categories. Furthermore, these fill rate differences (particularly for large orders) exceed the roughly one percentage point difference between the actual and hypothetical November Merrill Lynch orders used to validate the simulation (see Section 3).[21]

Fill rates suggest that the PHLX limit order trader enjoys an advantage over the NYSE trader for small OTQ limit orders. For large OTQ orders and for QI orders, the NYSE appears to offer higher fill rates than Philadelphia. Although the differences in fill rates are statistically significant, it is difficult to assess their economic importance. Filled orders economically benefit traders if security prices rise (fall) after buy (sell) orders execute. Conversely, traders experience regret if posttrade security prices move against them. To measure economic performance, we compute one-day returns for the package of limit orders sent to the PHLX and the identical package of hypothetical orders placed on the NYSE. The difference between these returns represents occasions where the stock's price touches the limit price but only one

[20] Although canceled actual Philadelphia orders are unfilled, we give hypothetical NYSE orders an opportunity to fill. If the NYSE order fills before the PHLX order cancels, then the NYSE order is considered filled. If the NYSE order is unfilled when the Philadelphia order cancels, then we also cancel the NYSE order. Thus, although the PHLX's fill rate for canceled orders is zero by definition, the NYSE's fill rate for canceled Philadelphia orders may be positive.

[21] Substantially more hypothetical QI orders fill because we observe a trade price that is inferior to our limit price in this simulation than in Section 3's test of the methodology. This is probably because an actual NYSE limit order exists in our test, but not in the simulation. We presume that the QI limit order executes when we observe sufficient trading at the limit price (classified as a "shares" trade) or worse (classified as a "trade" trade) on the NYSE. When the actual limit order exists, it is more likely that we observe an execution at the limit price. This distinction affects the classification of that execution, not if/when the order fills.

Table 6
Execution quality statistics for a sample of Philadelphia Stock Exchange limit orders and identical hypothetical New York Stock Exchange limit orders

Panel A: Fill rates

Order size (shares)	Limit order type	Order pairs	Fill rate (%) NYSE	PHLX	Difference	Hypothetical fills by (%) Shares	Trade	Quote
All	QI	9,558	91.63	89.40	−2.23***	81.53	17.84	0.64
All	OTQ	20,667	70.69	77.07	6.38***	85.33	3.52	11.15
100–599	QI	7,026	93.45	92.03	−1.42***	82.38	17.06	0.56
100–599	OTQ	13,988	74.11	84.83	10.72***	85.98	3.38	10.64
>599	QI	2,532	86.57	82.11	−4.46***	79.96	19.17	0.87
>599	OTQ	6,679	63.53	60.85	−2.68***	83.71	3.89	12.40

Panel B: Economic performance

Order size (shares)	Limit order type	PHLX fills, NYSE does not: Orders	Means return per share ($)	NYSE fills, PHLX does not: Orders	Mean return per share ($)	Total return difference ($)	Per share difference ($)
All	QI	119	0.0951	332	0.0198	1,402.94	0.0003
All	OTQ	1,880	0.1380	561	0.0507	73,509.38	0.0053
100–599	QI	81	0.1974	181	0.0912	185.44	0.0001
100–599	OTQ	1,648	0.1350	149	0.0903	60,767.50	0.0160
>599	QI	38	0.0493	151	0.0055	1,217.50	0.0003
>599	OTQ	232	0.1429	412	0.0480	12,741.88	0.0013

Panel C: Executive times

Order size (shares)	Limit order type	Filled order pairs	Mean time (minutes) NYSE	PHLX	Difference	Media time (minutes) NYSE	PHLX	Percent NYSE fills last (%)
All	QI	8,426	13.36	15.71	2.35***	3.15	4.28	21.75***
All	OTQ	14,049	40.50	32.34	−8.16***	14.97	10.70	38.73***
100–599	QI	6,385	12.00	14.01	2.01***	2.82	4.08	20.30***
100–599	OTQ	10,218	38.90	25.65	−13.25***	13.90	8.42	46.13***
>599	QI	2,041	17.62	21.03	3.41***	4.50	5.15	26.31***
>599	OTQ	3,831	44.78	50.17	5.39***	18.57	22.18	18.98***

The Philadelphia Stock Exchange (PHLX) data are collected from November 1, 1995, through January 31, 1996. QI indicates quote improving, i.e., an order with a limit price between the existing New York Stock Exchange (NYSE) bid and offer prices. OTQ indicates on the quote, i.e., a buy (sell) order with the limit price equal to the NYSE bid (offer) price. Overall, 1,283 PHLX orders with limit prices OTQ (or worse) relative to the national best quoted price become QI relative to the NYSE quote (i.e., there are 1,283 more QI orders here compared to the PHLX column in Table 1). Accordingly there are 925 fewer Philadelphia OTQ orders relative to the national best quoted prices (in Table 1) than relative to the NYSE quotes. The fill rate is the percent of orders submitted that fully execute on the day submitted. The column labeled difference is the PHLX fill rate less the NYSE fill rate. Three asterisks indicate that the hypothetical NYSE fill rate differs from the actual PHLX fill rate at the .01 level using a two-tailed t-test across individual orders. Hypothetical orders in the "shares" column execute because the number of shares executing at the limit price after the submission time equals the depth associated with the limit price at the time the limit order is submitted plus the size of the order. Hypothetical orders in the "trade" ("quote") column execute because we observe a transaction (quote) price that is inferior to the limit price (order submission time quoted price). To compute return, we compare the end-of-day national best bid (offer) price with the transaction price for filled buy (sell) orders. Filled buy (sell) orders in which the price rises (falls) have positive return figures. The number reported is the mean share-weighted figure. Total return difference subtracts the total return available for the hypothetical NYSE execution from the total return available from the actual PHLX execution. The per share difference is this total difference from the adjacent column divided by the total number of shares in the sample category. The execution time is the time required to fully execute the order. It is computed only for completely filled orders. The difference column is the PHLX mean execution wait less the NYSE mean execution rate. Three asterisks indicate that the mean hypothetical NYSE execution wait time differs from the mean actual PHLX execution wait time using a two-tailed t-test at the .01 significance level. The order-pairs in which the actual PHLX order executes no later than the hypothetical NYSE order as a percent of the total orders in the indicated category. Asterisks denote that the indicated percentage differs from 50% at the .01 significance level using a chi-squared test.

exchange fills the order before the price moves away and provides an estimate of the net economic value of the routing decision. Dividing this net return by the *total* PHLX-sample shares in the order category yields an estimate of the per share incremental benefit of the higher PHLX fill rate to the limit order trader. If the orders filled only in Philadelphia are associated with small returns, then our measure suggests the higher PHLX fill rate has little economic significance.

Panel B reports the per share economic performance measures for order pairs in which only the actual or only the hypothetical order executes and the net total economic performance for each category. The differences in economic performance for the QI orders are relatively small; only three-hundredths of a cent per share on average. Overall, OTQ orders executing in Philadelphia but not New York provide a one-day return $73,509 greater than the orders executing on the NYSE but not the PHLX. This is about one-half cent for each share included in the 20,667 OTQ orders. For small OTQ orders, Philadelphia fills 1,648 orders that are unfilled on the NYSE. On average, each of these shares increases in value by $0.1350 by the end of the day. The NYSE fills 149 orders that the PHLX does not and each of these shares increase in value by an average of $0.0903 by day's end. A positive return signals a one-day opportunity loss to routing those orders to the exchange not filling the order. For the small OTQ orders, there is a $60,768 advantage to PHLX routing instead of New York routing. This represents 1.6 cents per share in the 13,988 small OTQ orders, a figure of the same order of magnitude as deep-discount brokerage commissions for these trades. For large OTQ orders, the fill rate advantage favors New York. The economic advantage, however, favors the PHLX. The final row in panel B reveals why. Although the NYSE fills 412 orders Philadelphia leaves unfilled, while the PHLX fills only 232 orders unfilled in New York, the 232 orders increase in value by $0.1429 per share by day's end versus a $0.0480 per share increase in value enjoyed by the 412 orders filled only on the NYSE. The additional per share return associated with the marginal PHLX fills counters the more frequent fills on the NYSE for large OTQ orders.

To assess execution waits, we examine the 22,475 orders executing on both venues in panel C. Mean execution wait times for QI orders are significantly shorter in New York. Although the overall difference in QI orders' median waits is just 68 seconds, the NYSE order executes last in only 21.75% of the order pairs. The overall mean (median) wait for OTQ orders is 32.34 (10.70) minutes on the PHLX and 40.50 (14.97) minutes on the NYSE. Filled Philadelphia OTQ orders for fewer than 600 shares have a mean (median) wait 13.25 (5.48) minutes shorter than the NYSE's. Average waits for the 3,831 filled OTQ orders of 600 or more shares favor New York. Both the small- and large-trade mean wait time differences are significant at the .01 level. Furthermore, although the Philadelphia order executes no later than the NYSE order in 46% of the small-order order pairs, that happens in only 19%

of the order pairs for 600 or more shares. These findings reinforce the claim that the PHLX Rule 229.10 enhances the Philadelphia's limit order execution performance relative to the NYSE's.

5.2 Summary

We find that small OTQ Philadelphia limit orders appear to obtain higher execution quality than they would have received had they been routed to the NYSE. Specifically, we find that actual PHLX orders execute more frequently and that executed PHLX orders fill faster than the matched hypothetical NYSE orders. Furthermore, the net one-day return associated with being routed to the PHLX rather than to New York is \$0.0160 per share for these orders. Relative to our fundamental research issue, this suggests that the limit order routing decision does matter—OTQ limit orders eligible for execution under PHLX Rule 229.10 obtain better execution quality in Philadelphia than we estimate they would have obtained had they been routed to the NYSE. Conversely, we find that large OTQ orders and QI orders sent to Philadelphia are filled less often and that filled orders take longer to fill than identical hypothetical NYSE orders. Despite the higher NYSE fill rate, the PHLX retains a small economic advantage over the NYSE for these orders. Combining all orders, we find that the Philadelphia fill rate is a statistically significant four percentage points higher and the one-day return is \$0.0039 per share greater on the PHLX than it would have been on the NYSE. This suggests that the PHLX-sample limit orders benefit from being routed to Philadelphia rather than New York. These results may not be too surprising on a regional exchange with rules designed to provide such an advantage. In the following sections we examine data from other regional exchanges to determine the generality of this conclusion.

6. Preferencing Dealers on the Cincinnati Stock Exchange

In 1976 the Cincinnati Stock Exchange deemphasized its physical trading floor and initiated an automated market with competing market makers. Currently it conducts trading electronically via the National Securities Trading System (NSTS). The NSTS provides communication, order delivery, execution, and reporting capabilities to dealers in many locations. Members can (but are not required to) display limit orders in a central book and use NSTS to automatically execute orders according to priority rules. Market orders execute with the designated dealer (the CSE specialist) based on price and time priority, unless the firm entering an order indicates that a particular dealer is to receive the order (i.e., the broker "preferences" a particular dealer). Preferencing weakens time priority, but requires the indicated dealer to provide a price no worse than the NBBO. Brokers may internalize orders by preferencing an affiliated dealer, for example, Olde Discount Brokerage may own a CSE dealer to whom they send orders. If the broker wishes to internalize the order, then the broker sends both sides of the trade to NSTS (the broker-member on one side and the associated market maker on the other). After clearing same-priced public orders in the central limit order book, the order fills.[22] The SEC (1997) reports that the Cincinnati Stock Exchange had 72 members, 13 designated dealers, and 7 preferencing dealers in mid-1997.[23]

Although the Cincinnati Stock Exchange does not attempt to provide its limit-order traders an explicit advantage over New York traders, CSE Rule 11.9(u) requires members to benchmark limit order executions to New York. This rule states that public agency limit orders must trade if the size on the relevant side of the submission time primary market quote is exhausted or if the primary market's price becomes more favorable to the investor than the limit price. Brokers monitor what would have happened if the order had gone to New York instead of Cincinnati and execute the order when it is due an execution on either exchange. To monitor New York, brokers use a methodology similar to that described in Section 3 or actually submit the limit order (or a small "marker" order at the same limit price) to New York. The Cincinnati order executes when it reaches the front of either exchange's queue.[24]

An example of this print protection might help. At 12:32:33 on January 20, 1997, a Cincinnati broker-dealer receives a day limit order to buy 200 shares of Chrysler (now Daimler-Chrysler) stock for \$34.125. The dealer places this order at the top of its internal limit order book, because the dealer has no other buy orders for Chrysler at \$34.125. At the same time, the dealer sends an order to buy 200 shares of Chrysler at \$34.125 to the NYSE, where the quote is \$34.125 bid for 35,000 shares and 30,000 shares offered at \$34.25. The customer's limit order executes when the dealer receives a market sell order of sufficient size or when the limit buy order executes in New York. The order executes on the first venue on which it is eligible instead of sitting in a single-server queue. In theory, fill rates can be no lower than what would happen in New York, but may be better. In this particular case, the Cincinnati dealer obtains a market sell order for 200 shares of Chrysler at 13:12:58. The dealer cancels the New York limit order and crosses the order on the NSTS. The execution occurs before the limit order moves to the front of the queue on the NYSE allowing the customer to receive a more timely fill than if the order had been routed to New York. If the New York queue had been exhausted first, then the Cincinnati dealer would have obtained the customer's shares on the NYSE.

[22] Thus, time priority would not be weakened if limit orders went into the central limit order book. Most limit orders, however, exist only in the dealer's private book to avoid interfering with anyone's ability to internalize.

[23] See Easley, Kiefer, and O'Hara (1996), Bloomfield and O'Hara (1998), Kandel and Marx (1999), and Battalio, Greene, and Jennings (1997) for further discussion and treatment of internalization and preferencing.

[24] Other regionals have similar rules. See, for example, Boston Stock Exchange Rules, Chapter II, Section 33, and Chicago Stock Exchange Article XX, Rule 37(a).

To determine if Cincinnati dealers appear to enforce this New York-or-better standard, we use audit trail order data for five preferencing CSE broker-dealers. We note when a Cincinnati broker-dealer receives a limit order in an NYSE-listed common stock that is QI or OTQ in New York and place an identical, hypothetical limit order simultaneously on the NYSE. We use our estimate of if/when the NYSE orders execute and the actual CSE execution data to compare the likelihood of execution, the timeliness of the execution, and orders' economic performance. Fill rate comparisons use all 5,713 QI and OTQ order pairs.[25] About 29% of the orders do not execute on both the Cincinnati and the NYSE and are dropped from our execution wait analysis, which examines 4,070 order pairs. To analyze economic performance, we focus on the 368 order pairs filling on only one venue.

6.1 Fill rates, economic performance, and time to execution: CSE versus NYSE

Table 7 provides fill rates and waits and economic performance metrics for the CSE sample. Panel A reports average fill rates for the sample dealers. Of the 1,643 qualifying orders not filling on both exchanges, there are 1,275 orders not filling on either exchange, 188 orders filling on the CSE but not on the NYSE, and 180 orders filling on the NYSE but not on the CSE. For OTQ orders, the Cincinnati fill rate is 71.48% and the New York rate is 70.45%. A paired t-test finds that the OTQ fill rate difference (either overall or for the small order size category) is significant at the .01 level. As in Philadelphia, the NYSE provides significantly higher fill rates than Cincinnati for QI orders. If we accurately estimate New York executions, then these results suggest that the CSE brokers fulfill the New York-or-better execution standard for OTQ orders but fall short for QI orders.

Panel B's economic performance measures reinforce the fill rate results. Cincinnati provides an economic disadvantage for QI orders and an advantage for OTQ orders. The higher NYSE fill rate for QI orders provides an economic performance measure favoring the NYSE by \$0.0134 per QI share, a figure of the same order of magnitude as deep discount commissions for orders of this size. Although the OTQ fill rate difference between the CSE and the NYSE is small, orders filling on the CSE but not on the NYSE tend to have larger per share gains for the trader than orders filling on the NYSE but not on the CSE. This results in a \$0.0071 per share advantage for OTQ Cincinnati orders. Overall, investors are better off with the CSE routing by \$18,248.38 (= \$27,718.88 − \$9,470.50) or \$0.0040 per share.

In panel C we report execution wait results. Mean and median wait times for filled QI limit orders are similar on the NYSE and the CSE. The mean

[25] As with the Philadelphia sample, we must classify the hypothetical orders with respect to the NYSE quoted prices. This results in a Cincinnati sample of 268 more QI orders and 191 fewer OTQ orders than documented in the CSE column of Table 1, where we classify orders relative to the NBBO.

Table 7
Execution quality statistics for a sample of Cincinnati Stock Exchange limit orders and identical hypothetical New York Stock Exchange limit orders

Panel A: Fill rates

Order size (shares)	Limit order type	Order pairs	Fill rate (%) NYSE	Fill rate (%) CSE	Fill rate (%) Difference	Hypothetical fills by (%) Shares	Hypothetical fills by (%) Trade	Hypothetical fills by (%) Quote
All	QI	948	94.20	89.87	−4.33***	85.78	13.77	0.45
All	OTQ	4,765	70.45	71.48	1.03***	85.37	3.81	10.81
100–1099	QI	820	95.00	90.61	−4.39***	86.26	13.22	0.51
100–1,099	OTQ	4,059	71.99	73.24	1.25***	85.42	4.00	10.57
>1,099	QI	128	89.06	85.16	−3.90*	82.46	17.54	0.00
>1,099	OTQ	706	61.61	61.33	−0.28	85.06	2.53	12.41

Panel B: Economic performance

Order size (shares)	Limit order type	CSE fills, NYSE does not: Orders	CSE fills, NYSE does not: Mean return per share ($)	NYSE fills, CSE does not: Orders	NYSE fills, CSE does not: Mean return per share ($)	Total return difference ($)	Per share difference ($)
All	QI	4	0.3426	45	0.2920	−9,470.50	−0.0134
All	OTQ	184	0.2747	135	0.0638	27,718.88	0.0071
100–1,099	QI	3	0.1458	39	0.0898	−1,986.25	−0.0054
100–1,099	OTQ	162	0.2868	111	0.1140	13,369.88	0.0067
>1,099	QI	1	0.50	6	0.7142	−7,484.25	0.0022
>1,099	OTQ	22	0.2603	24	0.0145	14,350.00	0.0076

Panel C: Execution times

Order size (shares)	Limit order type	Filled order pairs	Mean time (minutes) NYSE	Mean time (minutes) CSE	Mean time (minutes) Difference	Median time (minutes) NYSE	Median time (minutes) CSE	Percent NYSE fills last (%)
All	QI	848	3.69	4.08	0.39	0.58	0.75	33.25***
All	OTQ	3,222	24.38	21.12	−3.26***	7.32	5.98	34.54***
100–1,099	QI	740	3.70	4.08	0.38	0.55	0.72	33.78***
100–1,099	OTQ	2,811	23.49	19.68	−3.81***	7.15	5.55	29.63***
>1,099	QI	108	3.62	4.11	0.49	0.78	1.21	36.14***
>1,099	OTQ	411	30.48	30.97	0.49	9.22	9.57	23.60***

The Cincinnati Stock Exchange (CSE) data are collected from January 20 through January 24, 1996. QI represents quote improving, i.e., an order with a limit price between the New York Stock Exchange (NYSE) bid and offer prices. OTQ represents on-the-quote, i.e., a buy (sell) order's limit price equals the NYSE bid (offer) price. Overall, 268 CSE orders with limit prices OTQ (or worse) relative to the national best quoted price become QI relative to the NYSE quote (i.e., there are 268 more QI orders here compared to the CSE column in Table 1). Accordingly there are 191 fewer Cincinnati OTQ orders relative to the national best quoted prices (in Table 1) than relative to the NYSE quotes. The fill rate is the percent of orders submitted that fully execute on the day submitted. The NYSE order is a hypothetical order. The difference column is the CSE fill rate less the NYSE fill rate. One (three) asterisk(s) in the difference column indicates that the hypothetical NYSE fill rate differs from the actual CSE fill rate at the .10 (.01) level using a two-tailed t-test across individual orders. Hypothetical orders in the "shares" column execute because the number of shares executing at the limit price after the submission time equals the depth associated with the limit price at the time the limit order is submitted plus the size of the order. Hypothetical orders in the "trade" ("quote") column execute because we observe a transaction (quote) price that is inferior to the limit price (order-submission-time quoted price). To compute return, we compare the end-of-day national best bid (offer) price with the transaction price for filled buy (sell) orders. Filled buy (sell) orders in which the price rises (falls) have positive return figures. The number reported is the mean share-weighted figure. Total return difference subtracts the total return available from the hypothetical NYSE fills from the total return available from the actual CSE fills. The per share difference is this total difference from the adjacent column divided by the total number of shares in the sample category. The wait time is the time required to fully execute the order. It is computed only for completely filled orders. The difference column is the CSE mean execution wait less the NYSE mean execution wait. Three asterisks indicate that the mean hypothetical NYSE execution wait time differs from the mean actual CSE execution wait time using a two-tailed t-test at the .01 significance level. The order pairs in which the actual CSE order executes no later than the hypothetical NYSE order as a percent of the total orders in the indicated category. Asterisks denote that the indicated percentage is different from 50% at the .01 significance level using a chi-squared test.

wait for OTQ orders is significantly shorter on Cincinnati than is the estimated mean NYSE wait. Although the mean wait time for OTQ orders is less on the CSE than the NYSE, the NYSE order fills after the Cincinnati order in only 34.54% of the order pairs. This suggests that, although the NYSE order typically fills first, large differences in wait times tend to favor Cincinnati. Although not reported in Table 7, the Cincinnati order fills within one minute of the New York hypothetical order in more than 48% of the OTQ order pairs. This is consistent with the claim that CSE dealers monitor NYSE executions and execute Cincinnati orders in reaction to trades in New York.

6.2 Summary

In contrast to PHLX Rule 229.10, which appears to produce a higher fill rate for eligible OTQ orders than is available in New York, print protection rules on the Cincinnati Stock Exchange produce OTQ fill rates that are comparable to the NYSE rates. We estimate, however, that the OTQ orders filling only on the CSE are more valuable to traders than those filling only in New York by $0.0071 per share. In addition, average OTQ order execution waits are shorter on Cincinnati than we estimate they would have been on the NYSE. We find QI orders routed to the CSE fill less often than do identical hypothetical NYSE orders. For QI orders, the one-day return associated with NYSE routing rather than CSE routing is $0.0134 per share. Comparing the total package of actual limit orders with the package of identical hypothetical orders, we find that fill rates and execution waits are nearly identical. The one-day return for the limit order package is $0.0040 higher on the CSE than we estimate it would have been had the orders been routed to the NYSE. This suggests that traders might not object to the Cincinnati routing, but that the overall routing decision for these orders is less important than in the Philadelphia sample.

7. Merrill Lynch Limit Orders

Finally, we use our Merrill Lynch data to compare actual limit order execution quality on the Boston Stock Exchange and the Pacific Exchange to the execution quality of identical, hypothetical limit orders submitted to the NYSE. Because Merrill's rerouting decision leaves no regional exchange orders in our November sample period, this comparison uses our October Merrill data. As with the Philadelphia and the Cincinnati comparisons, the order's limit price must be such that the hypothetical NYSE order is either QI or OTQ for us to infer the order's standing in the NYSE queue.[26] Merrill customers submit 634 such day limit orders in the relevant securities during

[26] When reclassifying the Boston (Pacific) orders with respect to the NYSE quoted prices rather than the NBBO quotes, we have 10(8) more QI orders and 9(6) fewer OTQ orders than in Table 1. Recall also that the Merrill's regional exchange data had to-the-minute time stamps. This requires us to delete orders arriving during a minute in which the NYSE quote changes because we cannot classify the order as QI or OTQ.

our five-day October sample period; 277 in Boston and 357 to the Pacific. Of the orders submitted, 178 BSE order pairs and 258 PCX order pairs fill on both the regional exchange and in New York. Conversely, 151 orders fill on neither venue (72 Boston orders and 79 Pacific orders). Twenty-three orders fill on the regional but not on the NYSE (14 Boston and 9 Pacific) and 24 fill on the NYSE but not on the regional (13 Boston and 11 Pacific).

7.1 Fill rates, economic performance, and time to execution: BSE and PCX versus NYSE

Table 8 provides fill rates and waits and performance measures for the October Merrill sample. We find that the estimated NYSE fill rate for QI orders exceeds either regional rate. The hypothetical NYSE rate is significantly higher than the BSE at the .10 level. Boston provides a significantly (at the .10 level) higher OTQ fill rate than we estimate these orders would have obtained on the NYSE. The NYSE hypothetical fill rate for Pacific OTQ orders is identical to the actual fill rate. Panel B reports economic performance measures. None of the total return differences are large. Combining both regionals, there is a \$413 (= \$248.38 + \$165.00) regional advantage. This is \$0.0013 per share of combined order flow to the two venues during the October sample period. Overall, Boston traders appear to be economically better off when comparing the return from orders filling only in Boston to those filling only in New York. This advantage is from the OTQ orders. For the Pacific, the economic advantage from the QI orders is offset by the OTQ disadvantage. In all cases, the sample sizes for the individual exchange results are small. The wait times reported in panel C reinforce the notion that Boston provides trade quality similar to the NYSE for limit orders. Mean and median Boston wait times are shorter than estimated NYSE times, with the mean OTQ (QI) wait time difference being statistically significant at the .01 (.10) level. Furthermore, in 53% (42%) of the OTQ (QI) order pairs, the actual Boston order trades first. On average, we estimate that filled PCX orders wait longer for execution than they would had they been submitted to the NYSE and find that in fewer than 28% of all order pairs does the PCX order fill first.

7.2 Summary

We find no statistically reliable fill rate differences between the actual PCX orders and what we estimate would have occurred had these orders gone to the NYSE. As with the PHLX and the CSE, the estimated QI (OTQ) fill rate on the NYSE exceeds (falls below) the actual QI (OTQ) rate in Boston, but only at the .10 significance level. Although Boston has an economic advantage over the NYSE, it is less than the advantage in the PHLX and CSE data. The Pacific's net economic performance advantage is nearly zero. We do find that filled orders appear to execute faster in Boston and slower on the Pacific than they would have in New York. As with the Cincinnati

Table 8
Execution quality statistics for a sample of Merrill Lynch limit orders on the Boston and Pacific Stock Exchanges versus identical hypothetical New York Stock Exchange limit orders

Panel A: Fill rates

Order size (shares)	Limit order type	Order pairs	Fill rate (%)			Hypothetical fill by (%)		
			NYSE	Regional	Difference	Shares	Trade	Quote
Boston	QI	117	84.62	79.49	−5.13*	75.76	22.22	2.02
Boston	OTQ	160	57.50	61.88	4.38*	81.52	3.26	15.22
Pacific	QI	114	92.98	91.23	−1.75	81.13	18.87	0.00
Pacific	OTQ	243	67.08	67.08	0.00	82.99	0.06	16.95

Panel B: Economic performance

Order size (shares)	Limit order type	Regional fills, NYSE does not		NYSE fills, regional does not		Total return difference ($)	Per share difference ($)
		Orders	Mean return per share	Orders	Mean return per share (%)		
Both	QI	3	$0.0833	11	−0.0378	248.38	0.0024
Both	OTQ	20	0.0640	13	0.0752	165.00	0.0007
Boston	QI	2	0.0667	8	0.0507	−92.25	−0.0017
Boston	OTQ	12	0.1183	5	0.0395	501.25	0.0063
Pacific	QI	1	0.1250	3	−0.2091	340.63	0.0067
Pacific	OTQ	8	0.0159	8	0.0896	−336.25	−0.0024

Panel C: Execution times

Order size (shares)	Limit order type	Filled order pairs	Mean time (minutes)			Median time (minutes)		Percent NYSE fills last (%)
			NYSE	Regional	Difference	NYSE	Regional	
Boston	QI	91	18.72	13.36	−5.36*	8.67	6.00	41.76
Boston	OTQ	87	40.42	27.90	−12.52***	17.53	13.00	52.89
Pacific	QI	103	16.66	29.51	12.85*	6.63	5.00	37.86**
Pacific	OTQ	155	88.03	99.72	11.69***	32.32	45.00	20.66***

The Merrill Lynch data are collected from October 2 through October 6, 1995. QI represents quote improving, that is, a buy (sell) order with a limit price between the New York Stock Exchange bid and offer prices. OTQ represents on the quote, that is, a buy (sell) order with a limit price equal to the New York Stock Exchange bid (offer) price. Overall, 10(8) orders on the Boston Stock Exchange (Pacific Exchange) with limit prices OTQ or worse relative to the national best quoted price become QI relative to the NYSE quote. That is, there are 10(8) more QI here than in the Boston (Pacific) column of Table 1. Accordingly there are 9(6) fewer Boston (Pacific) OTQ orders relative to the national best quoted prices than relative to the NYSE quotes. The fill rate equals the number of orders fully filled on the day submitted divided by the total number of orders submitted (times 100). NYSE indicates the New York Stock Exchange. Regional represents the Boston Stock Exchange (BSE) or Pacific Exchange (PCX). The NYSE order is a hypothetical order. The Difference column is the regional-exchange fill rate less the NYSE rate. One asterisk indicates that the hypothetical NYSE fill rate differs from the BSE fill rate at the .10 level using a two-tailed t-test across individual orders. No differences between the NYSE and the PCX are statistically significant. Hypothetical orders in the "shares" column execute because the number of shares executing at the limit price after the submission time equals the depth associated with the limit price at the time the limit order is submitted plus the size of the order. Hypothetical orders in the "trade" ("quote") column execute because we observe a transaction (quote) price that is inferior to the limit price (order submission time quoted price). To compute return, we compare the end-of-day national best bid (offer) price with the transaction price for filled buy (sell) orders. Filled buy (sell) orders in which the price rises (falls) have positive return figures. The number reported is the mean share-weighted figure. Total return difference subtracts the total return available from the NYSE hypothetical executions from the total return available from the regional executions. The per share difference is this total difference from the adjacent column divided by the total number of shares in the sample category. The wait time is the time required to fully execute the order. It is computed only for completely filled orders. The difference column is the regional mean execution wait time less the NYSE mean execution wait time. One (three) asterisk(s) indicates that the difference between the mean hypothetical NYSE execution wait time and the actual BSE/PCX mean execution wait time is significant at the .10 (.01) level using a two-tailed t-test. The order pairs in which the actual regional order executes no later than the hypothetical NYSE order as a percent of the total orders in the indicated category. Two (three) asterisks denote that the indicated percentage differs from 50% at the .05 (.01) significance level using a chi-squared test.

data, these results suggest that limit order traders, on average, are nearly indifferent between regional and primary exchange routing of their orders.

It is interesting to contrast the results in this section of the article to the results of the natural experiment. First, we note that our simulation results contradict the conclusion of the event study with regard to the relative execution quality of OTQ orders on the Pacific Exchange. In the event study, we find that November's OTQ fill rate is significantly less than October's and conclude that the execution quality of these orders suffers when orders are routed exclusively to the NYSE. The event study, however, does not control for changes in market conditions or order submission strategies between October and November. For example, Table 1 shows that the NYSE cancellation rate rises from 15% in October to 21% in November. Although it is beyond this article's scope to model changes in cancellation frequencies, the greater than six percentage point difference in the cancellation rate between the two adjacent months suggests using caution when drawing precise inferences from our event study. Second, we find that the actual Merrill NYSE quote-improving order fill rate in October documented in Table 1 (93.42% averaged across both the Boston and Pacific sample stocks) exceeds Table 8's hypothetical NYSE rate (88.75% is an order-weighted average of 84.62% and 92.98%). Likewise, for OTQ orders, the actual NYSE fill rate in Table 1 (64.25%) slightly exceeds the hypothetical NYSE fill rate in Table 8 (63.28%). These differences are consistent with Merrill acting in the best interests of its limit order customers in the October sample period by sending orders to the regional when market conditions favored higher fill rates there than in New York.

8. Summary and Conclusions

Prior studies of limit order trade quality either take the order routing decision as given or examine only one venue (e.g., the NYSE). These approaches are useful to document what occurs in markets, but do not address the question of whether orders would receive better fills elsewhere. For example, the SEC (1997) shows that fill rates are higher on the regional stock exchanges. Based on this evidence, can we conclude that regionally routed limit orders would have received worse executions had they been routed to the NYSE? Unless we believe that the regionals and New York receive the same types of orders in similar market conditions, we cannot reach such a conclusion from the SEC study. We use an event study and simulation to investigate the quality of limit order executions across trading venues. Our analyses are designed to control market conditions and order submission strategies when comparing limit order execution quality between regional stock exchanges and the NYSE. If execution quality differs across venues, then brokers can strategically route limit orders to improve customers' execution quality.

Our event study uses the natural experiment created with Merrill Lynch's rerouting of order flow from the Boston and Pacific Exchanges to New York. We compare limit order execution quality in NYSE-listed securities before and after the switch. Overall fill rates fall slightly afterwards for both the Boston and Pacific samples. Although execution waits for filled orders fall when orders are rerouted to the NYSE, the lower NYSE overall fill rates suggest that regional exchanges offer competitive limit order executions. That Merrill's limit order customers seem no better off, on average, after rerouting the orders contrasts with evidence in Battalio, Greene, and Jennings (1998) that Merrill's market order customers seem to receive better trade prices on the NYSE.

Because the narrow scope of our natural experiment does not allow a general evaluation of relative execution quality, we also develop and verify the accuracy of a simulation process designed to project the outcomes of hypothetical NYSE limit orders. This methodology facilitates a comparison of limit order execution quality while holding market conditions and order submission strategies constant across venues. As such, the simulation results also can be used by brokers to meet the SEC's mandated "regular and rigorous" review of their routing decisions. We use the size of the submission time NYSE quote to estimate our hypothetical order's position in the limit order queue and monitor executions and quote revisions to determine when our order "executes." Results of the validation test using actual NYSE day limit orders suggests that our methodology works well for the two most popular types of limit orders. We use audit trail databases of limit orders submitted to the Boston, Cincinnati, Philadelphia, and Pacific Stock Exchanges to investigate whether limit orders submitted to these regional exchanges would have enjoyed higher quality executions if they had been submitted to the NYSE. For each qualifying regional limit order, we simultaneously "submit" an identical hypothetical NYSE limit order and compare actual regional fill rates, execution times, and economic performance measures to hypothetical NYSE orders.

Considering all orders with limit prices bettering or matching the existing relevant quoted price (our sample), we find little difference between the regional exchanges' execution quality and our estimate of what would have happened had the orders been sent to the NYSE except for those orders routed to the Philadelphia Stock Exchange. Philadelphia limit orders qualifying for special treatment under a PHLX rule designed to produce high-quality limit order executions obtain significantly better execution quality than that apparently available concurrently on the NYSE for identical orders. Thus for brokers choosing between routing *all* (quote-improving or on-the-quote) orders in a given stock to either the PHLX or the NYSE, the order routing decision matters. For the comparable decisions when choosing between the NYSE and the Cincinnati, Boston, or Pacific exchanges, the execution quality enjoyed by the average limit order is nearly invariant to the broker's stock-by-stock routing decision.

If brokers make order routing decisions conditional on the limit price's proximity to the quoted price, then our simulations reach stronger conclusions regarding relative execution quality. For orders with limit prices between the existing NYSE bid and offer, results suggest that NYSE-routing would have improved execution quality. Specifically, the NYSE appears to offer higher fill rates and shorter waits than the Boston, Cincinnati, and Philadelphia Stock Exchanges for these orders. Combining the data from all four sample regional exchanges, we find that these orders' regional fill rate of 89.35% is significantly less than the estimated NYSE fill rate of 91.79%. This is not surprising. One downside of NYSE routing is the length of the typical limit order execution queue. For these quote-improving orders, there is no queue so the more frequent arrival of sufficiently large, opposite-sided market orders on the NYSE allows it to outperform the regionals. The execution quality of buy (sell) orders with limit prices equal to the existing NYSE bid (offer) price seems to be better on the Boston, Cincinnati, and Philadelphia Stock Exchanges than we estimate it would have been had the orders been routed to the NYSE. Again, pooling the four samples, we find that the overall regional fill rate, 75.86%, significantly exceeds the 70.53% NYSE rate. Thus brokers routed the on-the-quote orders to the regional exchanges to the apparent benefit of their customers, but also sent quote-improving orders to the regionals when they apparently could have received better executions on the NYSE. To the extent that brokers can make routing decisions based on the relationship between the limit price and the current quoted price, it may be to the customer's advantage to do so.

When combined with extant market order execution quality findings, our results suggest that a brokers' decision about where to route customer orders is difficult. Extant work [e.g., Blume and Goldstein (1992), Petersen and Fialkowski (1994), Bessembinder and Kaufman (1997), and Battalio, Greene, and Jennings (1998)] finds that market-order traders typically pay more for liquidity on regional exchanges than on the NYSE. We conclude that regional exchange routing of limit orders frequently appears justified, on average, providing higher fill rates and better economic performance measures. Unless brokers are willing to send orders in the same stock to multiple trading venues, the evidence is unclear as to which single venue is best. Our analysis is conducted taking the current fragmented market structure as given and our conclusions regarding investor welfare are conditional on that market structure. Finally, all of these studies provide information about the average, not marginal order routing decision.

References

Angel, J., 1994a, "Limit Versus Market Orders," FINC-1377-01-293, Georgetown University.

Angel, J., 1994b, "Who Gets Price Improvement on the NYSE?," working paper, Georgetown University.

Bacidore, J., R. Battalio, and R. Jennings, 2000, "Depth Improvement and Adjusted Price Improvement on the NYSE," Working Paper 2000–04, New York Stock Exchange.

Bacidore, J., K. Ross, and G. Sofianos, 1999, "Quantifying Best Execution at the New York Stock Exchange: Market Orders," Working Paper 99-05, New York Stock Exchange.

Battalio, R., J. Greene, and R. Jennings, 1997, "Do Competing Specialists and Preferencing Dealers Affect Market Quality?," *Review of Financial Studies*, 10, 969–994.

Battalio, R., J. Greene, and R. Jennings, 1998, "Order Flow Distribution, Bid-Ask Spreads, and Liquidity Costs: Merrill Lynch's Decision to Cease Routinely Routing Orders to Regional Stock Exchanges," *Journal of Financial Intermediation*, 7, 338–358.

Battalio, R., B. Hatch, and R. Jennings, 2000, "All Else Equal?: A Multi-Dimensional Analysis of Retail, Market Order Execution Quality," working paper, Indiana University.

Bessembinder, H., and H. Kaufman, 1997, "A Cross-Exchange Comparison of Execution Costs and Information Flows for NYSE-listed Stocks," *Journal of Financial Economics*, 46, 293–319.

Bias, B., P. Hillion, and C. Spatt, 1995, "An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse," *Journal of Finance*, 50, 1655–1689.

Bloomfield, R., and M. O'Hara, 1998, "Does Order Preferencing Matter?," *Journal of Financial Economics*, 50, 3–38.

Blume, M., and M. Goldstein, 1992, "Differences in Execution Prices Among the NYSE, the Regionals, and the NASD," Working Paper 4-92, University of Pennsylvania.

Chakravarty, S., and C. Holden, 1995, "An Integrated Model of Market and Limit Orders," *Journal of Financial Intermediation*, 4, 213–241.

Dow Jones New Service, 2000, "NYSE Disciplines 2 Firms, 4 Individuals," Dow Jones News Wire, Dow Jones & Co., February 23.

Easley, D., N. Kiefer, and M. O'Hara, 1996, "Cream-Skimming or Profit-Sharing? The Curious Role of Purchased Order Flow," *Journal of Finance*, 51, 811–834.

Glosten, L., 1994, "Is the Electronic Open Limit Order Book Inevitable?," *Journal of Finance*, 49, 1127–1161.

Handa, P., and R. Schwartz, 1996, "Limit Order Trading," *Journal of Finance*, 51, 1835–1861.

Harris, L., 1994, "Optimal Dynamic Order Submission Strategies in Some Stylized Trading Problems," working paper, University of Southern California.

Harris, L., and J. Hasbrouck, 1996, "Market vs. Limit Orders: The SuperDOT Evidence on Order Submission Strategy," *Journal of Financial and Quantitative Analysis*, 31, 213–232.

Hasbrouck, J., G. Sofianos, and D. Sosebee, 1993, "New York Stock Exchange Systems Entry Procedures," NYSE Working Paper 93-01, New York Stock Exchange.

Kandel, E., and L. Marx, 1999, "Payments for Order Flow on NASDAQ," *Journal of Finance*, 54, 35–66.

Kavajecz, K., 1999, "A Specialist's Quoted Depth and the Limit Order Book," *Journal of Finance*, 54, 747–771.

Kumar, P., and D. Seppi, 1993, "Limit and Market Orders with Optimizing Traders," working paper, Carnegie-Mellon University.

Lee, C., 1993, "Market Integration and Price Execution for NYSE-Listed Securities," *Journal of Finance*, 48, 1009–1038.

Lo, A., C. MacKinlay, and J. Zhang, 1997, "Econometric Models of Limit-Order Executions," Working Paper no. LFE-1031-97, Massachusetts Institute of Technology.

Macey, J., and M. O'Hara, 1997, "The Law and Economics of Best Execution," *Journal of Financial Intermediation*, 6, 188–223.

McInish, T., and R. Wood, 1995, “Hidden Limit Orders on the NYSE,” *Journal of Portfolio Management*, 21, 19–26.

New York Stock Exchange, 2000, *Fact Book for the Year 1999*, New York Stock Exchange, New York, NY.

Parlour, C., 1998, “Price Dynamics in Limit Order Markets,” *Review of Financial Studies*, 11, 789–816.

Petersen, M., and D. Fialkowski, 1994, “Posted Versus Effective Spreads: Good Prices or Bad Quotes?,” *Journal of Financial Economics*, 35, 269–292.

Ready, M., 1999, “The Specialists’s Discretion: Stopped Orders and Price Improvement,” *Review of Financial Studies*, 12, 1075–1112.

Ross, K., J. Shapiro, and K. Smith, 1996, “Price Improvement of SuperDOT Market Orders on the NYSE,” Working Paper 96-02, New York Stock Exchange.

SEC, 1994, Release No. 34-34753, “Customer Limit Orders.”

SEC, 1997, *Report on the Practice of Preferencing*, SEC, Washington, DC.

Seppi, D., 1997, “Liquidity Provision with Limit Orders and a Strategic Specialist,” *Review of Financial Studies*, 10, 103–150.

Sofianos, G., and I. Werner, 1997, “The Trades of NYSE Floor Brokers,” Working Paper 97-04, New York Stock Exchange.

Wall Street Journal, 1995, “Merrill Shifts on Small Investor’ Orders,” October 20, p. C1.