FINRA recently informed its member brokers that it would be conducting a review of their order routing practices. While I think that such an analysis is overdue, it should cause a great deal of consternation among broker dealers. In addition to the decision-making that goes into firms Smart Order Routers (SORs), the FINRA review is likely to dig into firm’s best execution policies to ensure that they properly evaluate routing choices. Importantly, this means that decisions made by the quantitative developers or traders at firms will need to be documented for best execution review purposes. Firms that use the algorithms and SORs of third party brokers will not be immune; in their case, they need to document how they evaluate both the choice of third parties and how they choose the algorithms and routing parameters.
So, What is FINRA Looking At?

they supply to that third party.

FINRA plans to focus on three questions, which all require that brokers have a quantitative best execution policy to answer them. While they are inter-related, each question looks at broker routing practices with a different angle, and will require targeted analysis, depending on the type of client and type of order received. Before explaining these, here are the specific questions, direct from the FINRA announcement:

1. How does the Firm quantify the benefits, if any, to [FIRM] customers from the Firm’s receipt of order routing inducements, such as payment for order flow and maker-taker rebates? Provide analytical or other evidence of such quantified benefits.

2. Describe how [FIRM] fulfills the Firm’s duty of best execution and quantifies the benefits, if any, to its customers when routing orders of a particular type to a market center with transaction costs for that order type that are materially higher than the transaction costs for the same order type on other market centers.

3. Describe how [FIRM] manages the conflict of interest that exists between the Firm’s duty of best execution to customers and the Firm’s own financial interests in situations where the Firm routes customer orders to market centers that pay order routing inducements, such as payment for order flow and maker-taker rebates, or internalizes customer orders (e.g. routing customer orders to an affiliated over-the-counter market maker or alternative trading system in which the Firm has a financial interest).

These questions can be deconstructed into three inter-related topics related to order routing: inducements, cost-based decision-making, & conflicts-of-interest. These are somewhat different for retail as opposed to institutional orders, but this commentary will focus on institutional clients “not held” orders. There are two reasons for this focus:
1. Relative need for analysis -- Retail “held” orders have been subject to best execution reviews and detailed analysis for many years already, and institutional “directed” orders do not involve decision-making by the broker.

2. Relative complexity of analysis – Most institutional orders are not routed to a single market-center for execution as one order, but are typically broken up into child orders which are routed over time. The exceptions to this are when a block trade is completed for the entire order, or when a broker routes a client order to another broker in its totality. In the first case, such orders are extremely easy to justify based either on client discretion or block liquidity availability. In the second case, however, the routing firm is going to need to defend how they analyze the detailed routing decisions made by the broker they used.

To get back to the FINRA questions, while they are related, they can be differentiated as follows:

**Inducements:** This area is predominantly focused on the question: How did the broker choose HOW and WHERE to post passive orders? This is because, on average, posting passive orders earns rebates from exchanges, with a larger rebate paid to displayed orders.

**Cost Based Decisions:** This area is focused mainly on the question: How did the broker choose how and where to route aggressive, liquidity-taking orders? This is because, on average, liquidity taking orders are charged fees from exchanges, with larger costs for accessing displayed orders on the primary listing exchanges.

**Conflicts of Interest:** This area is focused on the question of how the broker chose when to route passive vis a vis aggressive, liquidity taking orders? This question also permeates the two questions above, particularly when the option of posting orders on lower or non-rebate paying markets or taking liquidity on cheaper venues is available.

Let’s examine each question in turn, and describe the analysis that firms will need to provide FINRA when they start their review. At the risk of being redundant, even if the broker dealer has thought through all the logic and used data in their research, all of the points made below should be incorporated into the official best execution review process and be available for examiners.
To answer the question of how did the broker choose HOW and WHERE to post passive orders, firms will need to produce statistics that show fill rates and opportunity costs associated with the venues they used, alongside the rebates or fees of such venues. They will also likely need to document the data they relied upon to determine why not to route to venues that they did not use. As a practical matter, this breaks down into two types of analysis:

- **Order analysis**, meaning statistical measures of how orders on each venue compare based on two metrics:
  - Fill rates in context -- To place fill rates in context, this analysis must compare the fill rates of orders to the subsequent trades in the market to properly determine if the order could have been filled.
  - Opportunity costs for unfilled orders -- This measures stock price movement after order placement until the time the order is cancelled, where the cost is equal to that movement on the unfilled quantity.

- **Execution / market quality analysis** -- This is not a cost analysis, per se, but rather the type of supporting data used to justify routing decisions that includes two types of data:
  - Market quote quality statistics, obtained by analyzing public market data. Understanding which exchanges are more likely to be at the NBBO, set the NBBO, and what size they display at the NBBO are all part of algorithmic decision-making, and should therefore be available in best execution analysis.
  - Venue analysis based on the trades the broker has done. This would include measuring “toxicity,” which is defined as the post trade movement when orders are filled. This type of analysis helps to compare venue appropriateness for posting orders.
While SOR parameters might include explicit costs as an option, the main question is the same as with passive orders: fill rates and opportunity cost. For SORs, this translates directly into two key choices to measure: Routing Strategy and Routing Priority.

Routing Strategy -- Broadly speaking, there are two main types of routing strategy, which are each appropriate in different circumstances: Parallel and Sequential routing. Parallel routing refers to a strategy where the entire order is broken up into child orders that match what is displayed at individual markets at prices up to the limit price on the order and all are delivered to the markets displaying quotes at those prices simultaneously. Sequential routing, however, refers to strategies where the entire order is sent to one market at a time, often probing for midpoint liquidity or other price improvement, while also seeking more liquidity than is displayed by that market.

Note that understanding which strategy is best, depends on the market conditions immediately prior to routing. To explain why this is true, consider the following three scenarios, where, in all cases, the order being routed is to buy 5000 shares of XYZ:

- There are 6000 shares offered of XYZ across 4 different exchanges at the NBBO
- There are 200 shares offered of XYZ at 1 exchange at the NBBO
- There are 250,000 shares offered of XYZ across 9 different exchanges at the NBBO

In the first scenario, it is almost certain that a parallel routing strategy would be superior. If a broker used a sequential strategy, unless they found excess liquidity to fill the order, the pattern of trading created by trading portions of the order, would likely cause some or all the offers to vanish. (This could happen if competing algorithms saw the pattern and traded against the quotes before the sequential router could act, or simply because market makers saw the pattern and moved their quotes higher.) The most common metric for this scenario is to compare the actual effective spread achieved by the router to the quoted spread at the time the router sent its first order. That is fair, since there was sufficient liquidity available to complete the order at that time.
In the second scenario, however, it is not clear what strategy is best. A parallel strategy might well be too aggressive, pushing through multiple price levels simultaneously, but, it depends on the available liquidity at the next price level or two. In this scenario, the router would more likely probe sequentially for liquidity against dark pools and exchanges and might employ conditional orders or other liquidity seeking strategies as well. In this case, however, comparing the effective spread achieved by the router to the quoted spread is not fair. That is because the quoted spread did not reflect the fact that there was only a small fraction of the necessary liquidity available. To measure this scenario, a model such as ViableMkts Predicted Spread at Size (PSS) would be appropriate, as that is an impartial model which is designed to take available liquidity into account.

In the third scenario, parallel routing is almost certainly inappropriate, since a large amount of stock displayed at the NBBO is highly correlated to the potential availability of midpoint or price improved liquidity. In such a case, a sequential probing strategy is likely called for. To measure this scenario, the quoted spread could be used, but a model such as the PSS would be better, as it would capture the likely availability of stock inside of the quoted spread.

Routing Priority – In both types of routing strategies, there are choices that the routing firm makes about how venues are accessed. In parallel strategies, there is often more liquidity or insufficient liquidity available to complete the order. In the case of more liquidity than is needed, the SOR needs to determine which venues have priority, presumably based on the combination of predicted fill rate and potential price improvement. In practice, many SORs use relative explicit costs to make this determination, so the routing firm will need to show that such a choice did not result in inferior execution. In the case of insufficient liquidity available, the SOR needs to determine what venues to send the additional quantity to, which could be one of the exchanges or a dark pool.

In sequential strategies, the SOR needs to sequence the venues it accesses. It would be appropriate to separately evaluate midpoint probing from the NBBO orders, and it would also be smart to separate orders seeking liquidity from dark venues from exchanges. That said, it is important to look at sequential routing holistically, since that is the only way a firm can determine if they are “missing” liquidity at the NBBO when utilizing that strategy. To do so, the firm must be able to evaluate the order sent to the SOR against the NBBO and compare its aggregated effective spread and opportunity cost against the quoted spread or a modeled spread such as the previously mentioned PSS.
Can Conflicts of Interest Be Controlled?

While choices of specific venues to post passive orders and SOR strategies have potential conflicts of interest, arguably, the most conflict to be analyzed is whether the amount of passive, rebate receiving orders is justified. As I previously wrote in “Never Mind, These Are Not The Orders You Are Looking For”, on average aggressive orders provide better performance, when fees are not considered. Therefore, brokers will need to justify strategies that post significant percentages of passive, displayed orders.

To do so, they will need to document statistics such as the aggregate fill rates and opportunity cost associated with the unfilled orders, and not rely on “spread capture” statistics that ignore such costs. In situations where orders are posted away from the market to ensure participation when the market moves, those should be grouped separately, as those are routed for a different reason. (Note –the current 605 statistics already make the distinction between “at-the-quote,” “away-from-the-quote,” and “inside-the-quote” orders, and firms looking to be prepared for FINRA should do the same).

This type of analysis could be used to explain why firms route orders that earn rebates or route for lower cost and how important (or unimportant) the rebate itself is to the decision. It is certainly valid to argue that the client requested lower explicit commissions, so the key to a best execution justification is to be able to prove that the performance “loss” on a gross basis, was not greater than the reduction in commissions. That way, the broker can prove that the client was not harmed when everything is considered.
If I were at a broker dealer that used third party algorithms, I would be extremely nervous about the impending FINRA review.

I still have vivid memories of running an algorithm business at Citi’s Lava subsidiary, which sold algorithmic execution services to broker dealers, and remember quite clearly how cost conscious those brokers were. “My client is paying me the same commission, no matter what algorithm I use, so why should I pay more to you” was something I heard many times. I tried a couple of times to offer up “best execution” or “more liquidity for your clients” as an answer, but that was met by laughter from prospects, and anger from existing clients. I suppose it is possible that brokers are more concerned about their client’s performance today, compared to 9 years ago, but I doubt it. The bottom line is that ALL brokers are going to have to defend their choices of providers and the types of algorithms and routing strategies they use. Without a clearly articulated best execution policy, access to all the orders routed by the third-party providers and analysis of that routing, this will be impossible.