Experiments in High-Frequency Trading: Testing the Frequent Batch Auction

Eric M. Aldrich[*]
Department of Economics
University of California, Santa Cruz

Kristian López Vargas[†]
Department of Economics
University of California, Santa Cruz

March 26, 2018

Abstract

Using laboratory experiments, we compare two leading financial market formats in the presence of high-frequency trading (HFT): the Continuous Double Auction (CDA), also known as the continuous limit order book, which organizes trade in the majority of equities, futures and currency exchanges around the world; and the Frequent Batch Auction (FBA), which gives equal time priority to orders received within a short batching period. Our evidence suggests that, relative to the CDA, the FBA (1) reduces predatory trading behavior, (2) disincentivizes investment in low-latency messaging technology, and (3) results in lower transaction costs. Further, volatility in minimum spreads and in liquidity is higher in CDA compared to the FBA. Finally, we examine transitory, off-equilibrium behavior. In the CDA, transitory changes in the environment affect market dynamics substantially more than in the FBA.

Keywords: Market design, Auctions, high-frequency trading, Continuous Double Auction, Frequent Batch Auction.

JEL Classification: C91, D44, D47, D53, G12, G14.

Acknowledgements: The Center for Analytical Finance (CAFIN) at the University of California, Santa Cruz provided seed funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 741409). This paper is part of a larger, joint project with Dan Friedman, Axel Ockenfels, Peter Cramton. We thank them for invaluable feedback and Darrel Hoy and David Malec for developing the remote exchanges. We thank conference and seminar audiences at: the 2016 SEF meeting in Tucson, the 2017 SEF meeting in Nice (in particular, Jürgen Huber and Peter Bossaerts), the 2017 ESA meeting in Richmond, ICESI University and the UCSC-Econ brown bag seminar. We also thank Morgan Grant, Jason Vranek and Daniel Thurau for excellent programming work at the UCSC LEEPS Laboratory. The results reflect the authors' view, the ERC is not responsible for any use that may be made of the information it contains.

[*] Email: ealdrich@ucsc.edu.

[†] Email: kristian@ucsc.edu.

1 Introduction

Telecommunications technology has transformed financial markets in the last decade. Order submission and execution times at major exchanges has declined from seconds to microseconds (1 millionth of a second). As a consequence, traders are rewarded for reacting quickly to information, resulting in a new market participant: high-frequency trading (HFT) firms. HFT firms use computerized strategies to transact large volumes in fractions of a second and now account for more than half of transactions at major exchanges worldwide. Proponents claim that HFT has improved market liquidity (the ease with which trades can occur) and reduced transaction costs. Opponents argue that the multi-billion-dollar cost of HFT infrastructure is ultimately borne by investors, its liquidity is illusory, and it is a destabilizing force in financial markets.

Most academic papers studying HFT use proprietary data with trader identification, and are able to classify accounts as either aggressive (primarily liquidity consuming) or passive (primarily liquidity providing). Passive accounts are almost uniformly associated with improved market performance. Hagströmer and Nordén [2013], Malinova et al. [2014], Brogaard et al. [2015], Jovanovic and Menkveld [2015] and Menkveld and Zoican [2015] are examples of papers that find such positive effects. The effects of aggressive HFT, however, are mixed: it is generally associated with informed price impact over short horizons, increased adverse selection costs for other traders, increased short-term volatility, and higher trading costs for institutional and retail traders. Examples from this literature include Zhang and Riordan [2011], Bershova and Rakhlin [2012], Breckenfelder [2013], Brogaard et al. [2014], Hasbrouck and Saar [2013], Hendershott and Riordan [2013], Hirschey [2013], Baldauf and Mollner [2015a,b], Brogaard and Garriott [2015] and Menkveld and Zoican [2015].

Existing data are insufficient to resolve the controversy regarding HFT, as these data are almost exclusively collected from markets operating a single format: the continuous double auction (CDA). Without adequate counterfactuals, alternative market designs cannot be scientifically studied. Despite this, policy makers worldwide are already taking actions intended to discourage HFT. For example, in 2016, the U.S. Securities and Exchange Commission approved the Investors Exchange (IEX) to operate as a public securities exchange. A primary goal of the IEX market, a variant of the CDA, is to reduce potential advantages of HFT firms. In addition to the IEX, several other market formats have been proposed as alternatives to the CDA. These include the frequent batch auction (Budish et al. [2015], Budish et al. [2014]) and the fully continuous exchange [Kyle and Lee, 2017]. However, no scientific evidence exists on the relative performance of such market institutions.

In this paper, we use laboratory experiments to compare two leading financial market formats in the presence of high-frequency trading: the CDA and the frequent batch auction (FBA). The CDA (also known as the continuous limit order book) organizes trade in the majority of equities, futures and currency exchanges around the world. In this format, traders make publicly committed offers to buy and sell assets and are able to accept others' offers at any moment of time. In addition, traditional CDA markets employ a price-time priority system, which first ranks orders by price and then by the time they are received at the exchange (within price level). Because trading is continuous in time and orders are ranked by their submission time, communication speed is crucial in the CDA. Traders who can quickly react to new information have a substantial advantage over slower traders. This generates competition for speed technology, which is potentially socially inefficient. The FBA, on the other hand, does not allow trading to occur continuously. Instead, bids and asks are collected (batched) over discrete time intervals and call auctions are conducted at the end of each batching period. FBA therefore gives equal priority to orders received within a batching period, reducing the advantage of fast communication technology. Competition for speed becomes much less relevant than in the CDA and competition on price regains primacy.

The environment for our laboratory experiments is taken from the theoretical model of Budish et al. [2015] (hereafter referred to as BCS), where a single asset is traded on a single exchange. Traders submit commitments to buy or sell the asset in the form of *limit orders* and *market orders*. Two exogenous processes generate incentives to trade in this environment: changes in the publicly-observed fundamental value of the asset and the arrival of market orders from noise traders (*investors*) at random times. Although the two market formats price transactions differently, both assume that purchases (sales) are simultaneously liquidated (purchased back) at the fundamental value, allowing traders to book profits or losses instantaneously. Human participants (acting as traders) choose among three broad strategies to earn real profits: to act as *market makers*, to act as *snipers*, or to not participate in the market. Their choices can be continuously revised throughout the experiment. Market makers and snipers may subscribe to a technology that reduces messaging latency to the exchange for a pre-specified flow cost. Additionally, market makers choose a spread around the asset value at which they post bids to buy (below the value) and offers to sell (above the value). Market makers earn profit when the exchange matches them with a counterparty. Snipers attempt to exploit temporarily mispriced maker orders (stale quotes) by transacting with them at the time of a jump in the asset value.

To emulate modern financial markets, we develop an electronic architecture in which information and trading occur at millisecond time granularity. Specifically, for each format we deploy remote exchanges in an Amazon data center and utilize the Nasdaq OUCH protocol for messaging. Human subjects make high-level strategic decisions (outlined above) by interacting with a computer interface in the laboratory. Their decisions are encoded into algorithms which act on their behalf by communicating with the exchange as market conditions change. The arrival of exogenous information to the market via investor orders and changes in the fundamental value is designed to be representative of the time-scale in which similar information arrives in the market for a liquid asset, such as a the S&P 500 exchange traded fund. Thus, human subjects take the role of analysts at trading firms that design algorithmic strategies rather than the algorithms themselves.

The BCS model makes very sharp equilibrium predictions regarding the roles that subjects choose and their decisions to purchase fast communication technology. To understand the relationship of observed behavior to changes in the underlying environment, and hence predicted equilibria, we consider three treatments that vary the rate at which the asset value changes, the rate at which investors arrive, and the cost of fast communication technology. Regardless of treatment parameters, we find that subjects in the FBA market display substantially more passive liquidity, engage in less predatory behavior, and are less likely to purchase fast communication technology. Further, all FBA markets exhibit lower transactions costs, greater price efficiency, and less volatility. These results are all highly statistically significant and directionally support the predicted equilibria, although their values are typically somewhat attenuated relative to the precise equilibrium predictions.

Our work is as much a test of the BCS equilibrium model as a comparison of the CDA and FBA market formats. As such, our results are only relevant to real-world markets to the extent that the BCS model is a good characterization of actual trading behavior in those markets. However, testing the behavioral robustness of the BCS environment is interesting and useful in its own right. First, it is not clear, a priori, that human subjects can adequately learn, or that their behavior can converge in such a complicated and highly stochastic environment. Our results demonstrate that this is possible and can lead to useful insights. Additionally, the predicted equilibria are stylized and somewhat implausible due to their invariance over a wide range of the parameter space. Understanding behavioral divergences from these predictions will assist in the development of new theoretical models.

This paper contributes to the market design literature in finance and experimental finance. Related prior research on financial market design includes Roth and Xing [1994], who study the timing of transactions, Roth and Xing [1997], who study serial versus batch processing, Foucault [1999] and Roth and Ockenfels [2002], who introduce the idea of bid sniping, Ariely et al. [2005] who study how Internet auctions' ending rules shape the incentives for bid sniping, Du and Zhu [2017] and Fricke and Gerig [2015], who study the optimal frequency of double auctions, and Biais et al. [2014], who study "fast trading" and the externalities it generates. Relevant prior research on experimental finance include Friedman [1993], who reports on the first open tournament (for perishables) using a variant of the CDA, and Cason and Friedman [1996], Cason and Friedman [1997], Cason and Friedman [1999], and Cason and Friedman [2008], who compare variants of the CDA and call auctions (FBA) for perishables with independent private values.[1] We highlight that existing research on financial auctions focuses on environments that are not specifically relevant to the study of HFTs; to our knowledge, this paper reports the first experimental study of market makers in an environment with a common-value asset and noise traders as the primary source of profits.

The rest of the paper is organized as follows. Section 2 presents the experimental design and its implementation, Section 3 reports results of the experiments and Section 4 concludes. We collect discussions of model calibration and off-equilibrium behavior in Appendices A and C, respectively.

2 Experimental Design

We begin this section by describing the theoretical environment of the experiments and also providing detail on the two market formats that organize trade among participants. We then discuss the architecture of software and hardware used within the laboratory and for the remote exchange server. We conclude the section by describing implementation details related to treatments and the procedures of the experimental sessions.

2.1 Environment

Our laboratory environment is adapted from Budish et al. [2015] (BCS), where a single asset is traded on a single exchange. Traders express their willingness and commitment to buy or sell the asset by transmitting *limit orders* to the exchange. A limit order is a message comprised of four basic elements: (a) direction: buy (sometimes called a bid) or sell (sometimes called an ask or offer), (b) limit quantity (maximum number of units to buy or sell), (c) limit price (highest acceptable price for a bid, lowest acceptable price for an offer), and (d) time in force (indicating when the order should be canceled). A *market order* is a specialized limit order with the highest (lowest) possible limit price if it is a bid (offer). As a result, market orders transact immediately with any standing liquidity that expresses an opposing interest to buy or sell.

[1] General surveys on experimental research in financial markets can be found in Holt [1995], Sunder [1995], Friedman [2008], and Noussair and Tucker [2013].

Two exogenous processes generate incentives to trade in this environment: (1) the fundamental value of the asset, $V(t)$, which is publicly observed and evolves over time following a compound Poisson process with arrival rate λ_V and jump distribution F_V and (2) a population of *investors* (noise traders) that arrive at random times with Poisson rate λ_I, placing unit market orders to buy and sell with equal probability. Since investors exclusively use market orders, they transact immediately as long as there is countra-side interest. As described below, different market formats price transactions differently, but it is assumed that any purchase (sale) of the asset at any time is simultaneously liquidated (purchased) at the fundamental value, allowing traders to book profits or losses instantaneously.

The focus of the study is the behavior of N trading firms under differing market formats, and the outcomes this behavior generates. Human participants play the role of *trading firms* and, at any instant, can choose whether (a) to exit the market (*out*), or to participate either as (b) a *market maker* or (c) a *sniper*. In the latter two cases, traders also choose whether to invest in a technology that reduces round-trip messaging latency to the exchange from δ_{slow} to $\delta_{fast} < \delta_{slow}$ at a cost c_s per second. All traders are constrained to transact in unit shares of the asset.

Market makers are required to symmetrically post a buy order (*bid*) and sell order (*offer*) around the fundamental, $V(t)$. In practice, a maker chooses a spread, $s_i(t)$, which sets the price of her bid and offer to be, respectively, $V(t) - 0.5s_i(t)$ and $V(t) + 0.5s_i(t)$. Market makers earn profit when the exchange matches them with incoming investors, with the likelihood and magnitude of such transactions depending on the allocation mechanism of the market format.

When the fundamental value jumps from $V(t^-)$ to $V(t)$, a maker's orders will be temporarily mispriced at $V(t^-) \pm 0.5s$, where we assume that her choice of spread is fixed at $s_i(t) = s$. Immediately after the maker's algorithm learns about the new value $V(t)$, it submits a message to update her orders to $V(t) \pm 0.5s$. The update from $V(t^-) \pm 0.5s$ to $V(t) \pm 0.5s$ occurs at $t + \delta_{slow}$ by default, or earlier, at $t + \delta_{fast}$, if the maker subscribes to the fast communication technology at a price of c_s per second.

Snipers attempt to exploit stale quotes at the time of a jump in the fundamental value. When a value jump results in temporarily mispriced (stale) makers' orders, a sniper will attempt to transact with one of those stale orders to make a profit. Specifically, when the value jumps up (down), snipers try to quickly buy cheap from (sell high to) makers who have not yet updated their offers (bids). This is only possible when the jump is large enough and the market format allows it. As with makers, snipers' algorithms receive price information from the exchange and submit orders with default latency δ_{slow}, but can reduce their communication latency to δ_{fast} by paying c_s per second.

2.2 Market Formats

We now describe how the market formats we consider separately handle limit orders that are transmitted by traders.

2.2.1 Continuous Double Auction

Most modern financial exchanges implement a variant of the continuous double auction (CDA), also known as the continuous limit order book. This format is characterized by a *limit order book* that sorts limit orders by (1) price and (2) time received (at each price). Bids are sorted from highest to lowest price and offers are sorted from lowest to highest price. The highest bid and the lowest offer are referred to, respectively, as the *best bid* and *best offer*, and the difference between them is called the *spread*.

Traders (whether human or automated) enter, replace, modify and cancel orders at any moment they choose and the CDA processes each limit order as it arrives. If the limit price locks (equals) or crosses (is beyond) the best contra-side price – e.g., if a new bid arrives with limit price equal to or higher than the current best offer – then the limit order immediately transacts ("executes" or "fills") at that best contra-side price, and the transacted quantity is removed from the order book. On the other hand, if the price is no better than the current best same-side price, then the new order is added to the order book, behind other orders at the same price. The exchange breaks ties by randomly ordering messages received at the same price and time. Time priority in the CDA favors traders that can send, modify and cancel limit orders quickly, which results in all traders acquiring fast communication technology in the BCS equilibrium.

In this framework, market makers balance the profit generated by trading with investors (only if they post the best bid and offer) with the potential cost of being sniped at the time of value jumps and with the potential cost of purchasing fast communication technology. Under the CDA, BCS investors' sell orders transact with the highest maker bid and buy orders transact with the lowest maker offer. As a result, the maker with smallest spread, s, books profit $0.5s$ at a rate dictated by λ_I. However, at the moment of a sufficiently large positive (negative) change in the fundamental value, $J \equiv |\Delta V(t)| = |V(t) - V(t^-)| > 0.5s$, snipers will attempt to trade with the lowest (highest) maker offer (bid). A successful sniper earns profit $\Delta V(t) - 0.5s$ and the maker with smallest spread, s, takes a loss of the same size. The relative speed of other traders in the market dictates the probability that a maker is sniped when the fundamental value changes, as the exchange breaks ties by randomly ordering messages received at the same time. The equilibrium probability of such events is described below.

Snipers, on the other hand, balance the profits from sniping with the potential cost of fast communication technology. As with makers, their probability of earning profits is closely related to the relative speed of other traders in the market.

Given N trading firms participating in the market, Budish et al. [2015] show that the equilibrium of their model under the CDA format consists of a single market maker, $N-1$ snipers, and all traders purchasing fast communication technology. Their equilibrium is characterized by two zero-profit conditions. For the maker,

$$\lambda_I \cdot \frac{s}{2} - \lambda_V \cdot \Pr\left(J > \frac{s}{2}\right) \cdot E\left[J - \frac{s}{2} | J > \frac{s}{2}\right] \cdot \frac{N-1}{N} = c_s, \tag{1}$$

which states that the expected profits from trading with investors, $\lambda_I \cdot \frac{s}{2}$, less the expected losses to snipers, $\lambda_V \cdot \Pr\left(J > \frac{s}{2}\right) \cdot E\left[J - \frac{s}{2} | J > \frac{s}{2}\right] \cdot \frac{N-1}{N}$, must be equal to the cost of buying speed services. For the sniper,

$$\lambda_V \cdot \Pr\left(J > \frac{s}{2}\right) \cdot E\left[J - \frac{s}{2} | J > \frac{s}{2}\right] \cdot \frac{1}{N} = c_s, \tag{2}$$

which says that the profits from sniping, $\lambda_V \cdot \Pr\left(J > \frac{s}{2}\right) \cdot E\left[J - \frac{s}{2} | J > \frac{s}{2}\right] \cdot \frac{1}{N}$, must be equal to expenditure on speed services. Note that sniping profits are weighted by different fractions of N in Equations (1) and (2); this is a result of the fact that all traders are subject to identical communication latency δ_{fast}, causing the messages of all N players (the single maker and $N-1$ snipers) to be received by the exchange at the same time. With probability $\frac{N-1}{N}$, the maker loses $\lambda_V \cdot \Pr\left(J > \frac{s}{2}\right) \cdot E\left[J - \frac{s}{2} | J > \frac{s}{2}\right]$ to one of the snipers and with probability $\frac{1}{N}$ one of the snipers will earn the same amount.

Equations (1) and (2) endogenously determine both the maker's equilibrium spread, s^*, and the total number of trading firms, N^*. Budish et al. [2015] re-expresses the equilibrium as

$$\lambda_I \cdot \frac{s^*}{2} = \lambda_V \cdot \Pr\left(J > \frac{s^*}{2}\right) \cdot E\left[J - \frac{s^*}{2} | J > \frac{s^*}{2}\right] \tag{3}$$

$$\lambda_I \cdot \frac{s^*}{2} = N^* c_s. \tag{4}$$

Equation (3), which is the difference of Equations (1) and (2), determines s^* and Equation (4), which is the sum of Equation (1) and $N-1$ times Equation (2), determines N^* for a given s^*. Budish et al. [2015] interpret Equation (4) as showing that the cost of speed, purchased by all traders, is borne entirely by investors via transactions with market makers.

All traders purchase fast communication technology in the CDA equilibrium to either prevent severe loss (maker) or to prevent exclusion from profits (snipers). Abstaining from fast communication is not a profitable deviation: a slow maker increases his chances of getting sniped by $1/N$ and a slow sniper reduces her sniping chances to zero.

2.2.2 Frequent Batch Auction

In a Frequent Batch Auction (FBA), the trading day is divided into submission stages of equal length τ. These submission stages are referred to as *batching intervals* or *batches* and can be considered discrete time increments. Within a batch, traders communicate with the exchange in continuous, sealed-bid fashion in order to privately submit, modify and cancel limit orders, which are collected and held by the FBA matching engine.

At the conclusion of a batch, all orders are combined with unfilled orders from previous batches and the matching engine generates stair-step demand and supply curves from the aggregated bids and offers, respectively. If demand and supply do not intersect, no trade occurs and all orders carry over to the next batch, aside from those denoted as *immediate or cancel.* If demand and supply intersect, the market clears where supply equals demand, i.e., all infra-marginal bids and offers are executed at a uniform price p^* that clears the market. The FBA matching engine then publicly broadcasts information regarding executed trades, p^*, and the remaining order book.

Figure 1 diagrams the central features of an FBA batch. For a batch concluding at time t, traders subject to communication latency δ_i are not able to act on new information during time interval $(t - \delta_i, t]$ before the time t auction. For example, if the fundamental value of the asset changes in interval $(t - \delta_{slow}, t]$, slow makers will not have an opportunity to update their bids and offers and slow snipers will not have an opportunity to submit aggressive market orders. The same is true of fast makers and snipers for asset value changes in the interval $(t - \delta_{fast}, t]$. The difference $\delta = \delta_{fast} - \delta_{slow}$, in relation to the total batch length τ, represents the relative advantage of fast communication technology: as the ratio δ/τ declines, one expects the value of speed technology to diminish.

Under certain conditions, the BCS equilibrium under FBA consists of all traders choosing to act as market makers with zero spread, and none of them purchasing fast communication technology.

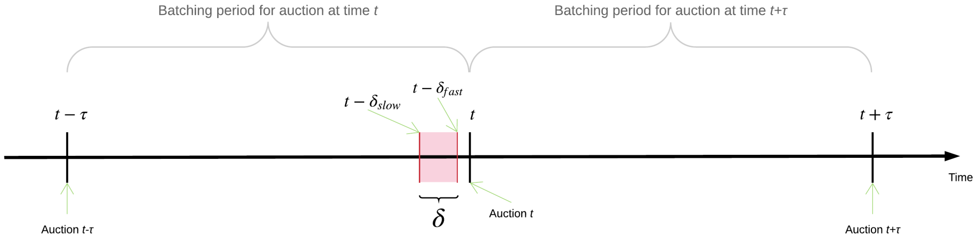


Figure 1: Timing in the FBA format (adapted from Budish et al. [2015]).

Specifically, Budish et al. [2015] show that if:

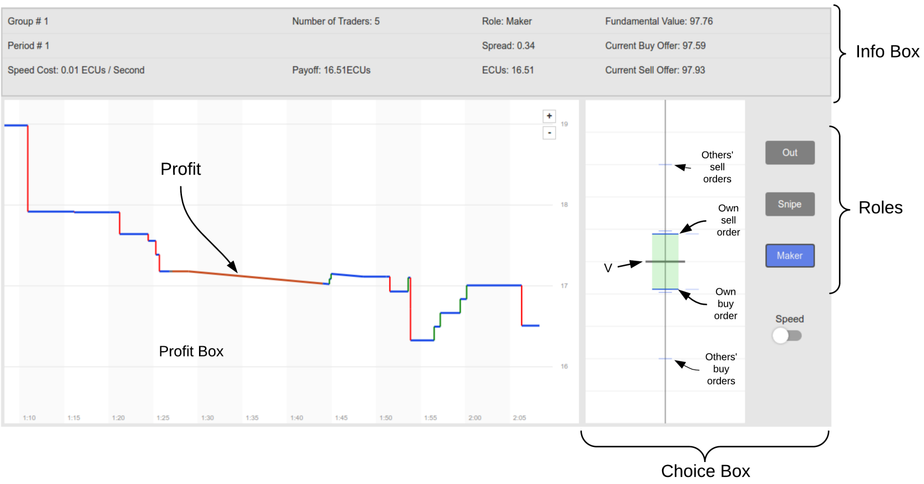
$$\frac{\delta}{\tau} \cdot \lambda_V E\left[J|J>0\right] < c_s \tag{5}$$

there is no incentive for traders to take the role of a fast sniper in the FBA. Intuitively, the left-hand-side of Equation (5) is the profit attributed to a fast sniper who successfully exploits a unit share posted by a slow market maker when a jump occurs in the interval δ: $\frac{\delta}{\tau} \cdot \lambda_V$ represents the probability of a value jump in the short interval δ and $E\left[J|J>0\right]$ represents the magnitude of the gain (when $s^* = 0$). Since there are no fast snipers in the market, the need for high-speed communication technology vanishes. Competition now focuses purely among makers on price (à la Bertrand) and dictates that all market makers undercut each other until $s^* = 0$ (all bids and offers are set to equal the observed fundamental value, V).

It is important to note that, as in the CDA, makers earn profits in the FBA by transacting with investors. However, since investors submit the best bids and offers within a batching period, the FBA auction often pairs their orders together as transactions. As a result, makers only transact with investors when the number of buying investors within a batch differs from the number of selling investors. Such transactions occur at the market clearing price, p^* which is weakly lower (higher) than the best bid (offer). Thus, for a given set of player strategies in the market, there are two reinforcing channels by which makers earn lower transaction profits in the FBA relative to the CDA: (1) fewer investor transactions and (2) transactions prices that are closer to the fundamental value of the asset. This reduction in profits is partially offset by smaller losses to sniping.

2.3 Architecture

Communication latency and the speed with which traders implement their strategies is a central feature of the BCS equilibrium. However, these latencies are technological constraints and not related to traders' physical limitations. Thus, tests of the BCS equilibrium should not be dependent on human reaction time. To meet this objective, human subjects in our experiment interact with a high-level computer interface that tunes an algorithmic strategy. Subjects' algorithms interact with a remote exchange server at millisecond time granularity and ensure that a particular strategy is properly implemented as market conditions change. For a fee (described below), participants may reduce their algorithm's messaging latency with the exchange.



Group # 1 | Number of Traders: 5 | Role: Maker | Fundamental Value: 97.76

Period # 1 | Spread: 0.34 | Current Buy Offer: 97.59

Speed Cost: 0.01 ECUs / Second | Payoff: 16.51ECUs | ECUs: 16.51 | Current Sell Offer: 97.93

Figure 2: CDA user interface in the BCS environment.

2.3.1 Laboratory

The laboratory component of the architecture is implemented using the Redwood 2 platform (`https://github.com/RedwoodAdmin/RedwoodFramework`). Redwood 2 is a Django-based experiment platform designed for games in continuous time. It is comprised of the following modules:

1. A graphical user interface that displays payoffs, recent history and current state of the market, and trader controls.

2. Individual algorithms (one per subject) that keep track of subjects' choice states and implement their strategies by automatically sending messages to the remote exchange (using Nasdaq's OUCH 4.2 protocol) as market conditions change.

3. A market manager that handles OUCH messages composed by subjects' algorithms and passes them on to the remote exchange. In return, the market manager receives confirmation messages from the exchange, which are forwarded to individual algorithms. The market manager also implements value jumps and investor arrivals from prespecified draws of the Poisson processes describe above.

The user interface for the CDA format is shown in Figure 2. The *info box* (top) displays basic information about the market state such as the current bid-offer spread, the number of active traders and the cost of speed c_s. Traders use the *choice box* (bottom right) to adjust choices at any moment during the trading period. To (re)enter as a market maker, a subject clicks the "Maker" button, or selects a spread (s_i) by clicking in the white area of the *choice box*. To (re)enter as a sniper, a subject clicks the "Snipe" button. By clicking the "Out" button, a subject cancels any limit orders, deactivates any algorithms trading on her behalf, and unsubscribes from speed services. The *profit box* (bottom left) displays subjects' accumulated profit at each moment of time. A subject's profit jumps if there is an investor arrival that executes against her buy or sell orders, if she snipes, or if she gets sniped.

The interface for the FBA format is essentially the same as in the CDA. The FBA choice box includes a time marker that indicates the length and the moment of the auction and the profit box includes markers that indicate the instant of each auction (see Appendix B for a screen shot of the FBA user interface).

2.3.2 Exchange

Our experimental exchange is patterned after the architecture of modern financial exchanges. In particular, it is comprised of two components: a *messaging server* and a *matching engine*. The messaging server follows the Nasdaq OUCH 4.2 specification, which strictly defines a set of incoming and outgoing private messages and their format (the order and byte length of each field within a message). The most common incoming messages (from participant to exchange) sent via OUCH are new order submissions, as well as updates and cancellations of existing orders. Outgoing OUCH messages (from exchange to participants) consist of private confirmations of incoming messages. In addition, the messaging server follows the Nasdaq ITCH 4.1 protocol for public broadcasts of market information.

The matching engine processes orders relayed from the messaging server, executes transactions according to a specified format, and reports the results to the messaging server. In both formats we consider, the matching engine maintains a limit order book – the set of unexecuted, actionable limit orders in the market at a given point in time.

Both components are hosted on an Amazon's Elastic Compute Cloud (EC2) infrastructure located in California, enabling tight control over physical accessibility and communication latencies.

2.4 Treatment Design

A treatment in our experiment consists of a combination of *market format* and a *market configuration*. We study two market formats (CDA and FBA) and three market configurations. In our baseline configuration (C1), we set $\lambda_I = 1/3$, $\lambda_V = 1/4$ and $c = 0.01$. As argued in Appendix A, these values are suggestive of actual dynamics observed in the S&P 500 exchange traded fund. Substituting these values in the BCS equilibrium Equations (3) and (4), results in $s^* = 0.324$ and $N^* = 5.4$ under CDA. Under FBA, $s^* = 0$ and $N^* = 6$ (since there are a total of 6 traders in a market).

For the remaining configurations, we consider variations in the relative jump and investor arrival intensities, as well as the cost of speed, in order to induce larger equilibrium spreads while maintaining the same endogenous population of traders, $N^* \approx 5.4$. In configuration 2 (C2), we use $\lambda_I = 1/5$, $\lambda_V = 1$ and $c = 0.0104$. This could be loosely interpreted as a volatile market, with many asset value changes (once per second) and few investor arrivals. In this case the CDA equilibrium spread is much higher: $s^* = 0.566$. Finally, in configuration 3 (C3) we consider $\lambda_I = 1/2$, $\lambda_V = 1$ and $c = 0.022$, which doubles the cost of speed technology while maintaining the rapidly moving market, but with relatively more investor arrivals. Under this parameterization, the CDA equilibrium spread is $s^* = 0.475$. As with C1, $s^* = 0$ and $N = 6$ under FBA for C2 and C3. Panel (a) of Table 1 shows the parameters for each market configuration. Equilibrium predictions for these configuration parameters are shown in Panel (b) of the same table.

In addition to the parameters above, we set $\delta_{slow} = 0.5$ seconds, $\delta_{fast} = 0.1$ seconds for both CDA and FBA in all configurations. We fix the length of a batching period to be $\tau = 3$ seconds, which means that the fraction of time that snipers can exploit slow makers under FBA is $\frac{\delta_{slow} - \delta_{fast}}{\tau} = \frac{0.4}{3} = 0.133$.

We used a between-subject design to collect data from four independent groups (markets) for each of the six treatments $\{CDA, FBA\} \times \{C1, C2, C3\}$. Each session consisted of 12 traders divided in two groups (markets) of six participants and a single treatment was implemented in a session. A session consisted of eight consecutive trading periods of four minutes, with fixed groups across all periods (partner matching). To obtain data for 24 markets (6 treatments × 4 groups), we conducted a total of 12 sessions. Within a trading period, participants were able to change strategies at any moment. However, as highlighted above, the study and architecture were designed to test traders' strategic behavior rather than human reaction times. In all configurations, $\Delta V(t) \sim \mathcal{N}(0, 0.5)$, as described in Appendix A.

2.5 Procedures

Sessions were conducted at the LEEPS Laboratory of the University of California, Santa Cruz. Recruitment was implemented through LEEPS' ORSEE instance (`econlab.ucsc.edu`; Greiner [2015]). Instructions were provided on the computer screen. After 15 to 20 minutes of reading instructions, a trial round of 90 seconds was launched during which subjects tested the interface with the understanding that their actions were not being recorded and would not comprise part of the formal experiment. Subjects were then given the option to publicly ask questions to the experimenters. The remainder of the session consisted of eight trading periods of four minutes. In between trading periods, subjects received a summary screen displaying the distribution and profits of different roles in the corresponding trading period. Each participant received 20 experimental currency units (ECUs) as initial endowment at the beginning of each trading period and final payments were based on the final wealth of a randomly chosen period, using an exchange rate of two ECUs per dollar, plus a $7 participation fee. Although ending a trading period with negative profits was possible (i.e., losing 20 ECUs in the period), this happened only in 1.3% of the cases and subjects never left the laboratory with less than $7. Members of a group (market) were not able to communicate with each other before, during or after the experiment, nor did they learn the identity or characteristics of other participants. Payments were implemented following standard procedures of confidentiality.

3 Results

We compare the CDA and FBA market institutions with the following metrics: (1) market liquidity, measured by the fraction of traders who choose to be *market makers*, which is a surrogate for the depth of the order book in this environment; (2) the prevalence of predatory behavior, measured by the fraction of traders who choose to be *snipers*; (3) the penetration of high-speed communication technology in the market, measured by the fraction of traders that choose to subscribe to this service; and (4) transaction costs for investors, measured by the minimum spread among market makers. We also study other auxiliary metrics associated with volatility (e.g., the standard deviation of the minimum spread) and informational efficiency (such as the root mean squared deviation between realized prices and fundamental value). Measures of allocative efficiency are less relevant for this paper because we are dealing with a common value asset.

Prior to reporting our experimental results, several observations are in order. First, one of the CDA sessions (configuration 3) was terminated after 7 (rather than 8) periods because the 1.5-hour session time limit was reached. This affected two of the four markets for that configuration. Second, in certain circumstances, order repricing for market makers resulted in maker-to-maker sniping events. Appendix C describes these transactions in more detail and shows that they do not affect the BCS equilibrium. We discovered maker-to-maker snipes during initial pilot sessions and decided not to eliminate them, since doing so would violate typical order book rules. However, we did not explicitly acknowledge this feature in the experiment instructions, instead allowing subjects to observe the consequences of these events on their profit and decision screens. Finally, after collecting our experimental data, we discovered that the FBA exchange did not randomize transaction allocation for orders received within the same batch and at the same price. Under FBA, time priority should not exist for orders received by the exchange during the same batching period; such priority should only exist for orders received in prior batches and which have not been filled. Although this is a violation of the FBA matching mechanism, it affected a very small fraction of orders (only those with identical prices within a single batch) and, at worst, provided an increased incentive to purchase fast communication services (which is not the equilibrium). Revising the system to allow for randomization of orders during auctions would potentially yield results that are even more congruent with predicted equilibria.

3.1 Subject Choices

Figure 3 depicts the four summary measures, averaged across all groups within a given treatment. The faint, solid lines represent average values computed at five-second intervals and bold, solid lines represent exponentially-weighted moving averages with ten-second half lives. Dashed lines represent equilibrium values under the BCS model. Each panel of the plot depicts the full time series for a session, concatenating the metrics across the eight four-minute periods, which are visually separated by alternating gray and white vertical bands. To smooth start-up effects within periods, in which players' default state is "Out", we do not include the first ten seconds of each period in the plot.

The blue lines in Figure 3 correspond to equilibrium and observed values under CDA. For the first three metrics, the BCS equilibrium values are constant across configurations: the percentage of makers is predicted to be $\frac{1}{N} * 100 = 100/6 = 16.67\%$, the percentage of snipers is predicted to be $\frac{5}{N} * 100 = 500/6 = 83.33\%$, and the percentage of players buying speed services should be 100%. For the final metric, equilibrium spread, the values for each configuration (reported in Section 2.4) are $s^* = 0.324, 0.566$ and 0.475 for configurations 1, 2 and 3, respectively. The green lines in the figure correspond to equilibrium and observed values under FBA. The predicted BCS equilibrium values for all four metrics are constant across configurations: 100% market makers, 0% snipers, 0% purchasing speed services and $s^* = 0$.

Qualitatively, the time series show a separation in the direction of the predicted equilibria: for all configurations (1) the fraction of makers is higher under FBA, (2) the fraction of snipers is lower under FBA, (3) the fraction of traders purchasing speed services is lower under FBA, and (4) the equilibrium spread is lower under FBA. In some cases, the data suggest mild evidence of learning during the first 2 or 3 periods, and in nearly all cases the time series show a clear statistical separation (to the extent that the five-second averages exhibit accurate variation in the data) after the initial learning period, if not before. In magnitude, few of the observed metrics achieve equilibrium values, with the consistent exception of the FBA spread. In this case, the observed values remained at or near the floor of 0.1 ECUs implemented in the lab interface.

Recall that relative to configuration 1, the market under configuration 2 experiences jumps in the fundamental value roughly four times more often, has 40% fewer investors, and has nearly identical cost of speed. Configuration 3, on the other hand, also has four times as many jumps in the fundamental value, but receives 1.5 times the number of investors, and has double the cost of speed. Thus,

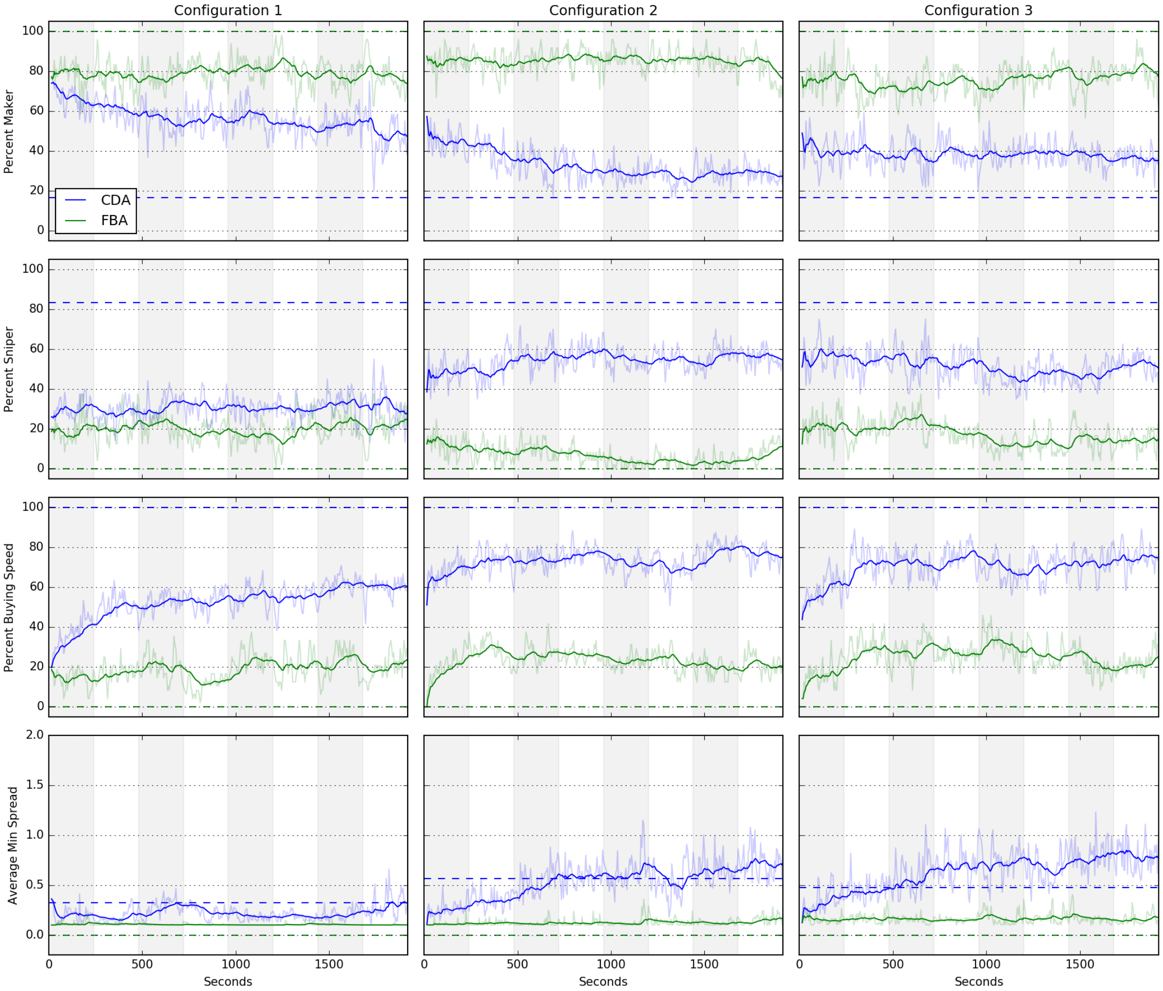


Figure 3: Time series of subjects' actions: strategies, speed subscriptions and minimum spread. In each box, we plot traders' strategies in CDA (blue lines) and FBA (green lines). Each column of the figure represents a market configuration and each row represents a variable related to the average of the outcome across groups. The dashed lines correspond to theoretical equilibrium values predicted by the BCS model. Faint solid lines correspond to observed time series sampled at five-second intervals and bold solid lines are their exponentially-weighted moving averages with a ten-second half life.

both configurations 2 and 3 can be regarded as volatile markets, which is costly to market makers, but configuration 3 partly compensates makers with additional income through investors, whereas configuration 2 provides even lower investor income, relative to the baseline calibration. In addition, under configuration 3, all subjects bear a higher cost of speed (makers and snipers).

Accordingly, the time series show that the fraction of market makers under CDA (FBA) configuration 2 is lowest (highest), corroborating that the costs of providing liquidity is highest in that configuration and that FBA offers the most protection in those circumstances. Similarly, the fraction of snipers and the fraction of traders purchasing speed technology under CDA configuration 2 is higher than CDA configuration 1, and weakly higher than CDA configuration 3. For FBA, the observed fraction of snipers and fraction of traders purchasing speed in configuration 2 are equal to or lower than those values in the other configurations. Generally speaking, the observations for configuration 3 lie between those of configurations 1 and 2, as would be anticipated from the high volatility coupled with somewhat higher investor income to makers (yet, despite the high cost of speed, which should have a greater effect under CDA). In each case, the observed spreads do a reasonable job of tracking predicted equilibria, although CDA configuration 3 is notably high.

Panel (b) of Table 1 reports summary statistics for the metrics outlined above, sampled at one-second intervals, averaged across both time and groups and within treatment. As with Figure 3, we exclude the first ten seconds of each four-minute period in order to eliminate start-up effects, and we also exclude the first two periods of each session, to account for learning.

Consistent with Figure 3, the values in panel (b) of Table 1 demonstrate that, relative to the CDA, under the FBA (1) more traders choose to act as makers, (2) fewer choose to act as snipers, (3) fewer choose to purchase speed services, and (4) minimum spreads are smaller. We provide a detailed discussion of these values in Section 3.3, in the context of a regression analysis which measures the difference in means across market formats.

3.2 Market Statistics

Panel (c) of Table 1 reports summary statistics that measure volatility of prices, volatility of strategy choices (changes of roles), pricing deviations from fundamental value, number of transactions, and trader profits. Specifically, the first row of Panel (c) provides a measure of transaction price volatility via the standard deviation of price differences. The data show that observed price differences are between five and 7.5 times more volatile in the CDA configurations, with price volatility higher (for both CDA and FBA) in configurations 2 and 3. In equilibrium, these values should be close to the standard deviation of jumps in the fundamental value, which was set to 0.5. The second and third rows of panel (c) report volatility measures related to traders' choices: the standard deviation of the minimum spread and the average number of changes in trading strategy (status) by subjects within a trading period. In equilibrium, the standard deviation of minimum spread should be zero since makers always choose the fixed equilibrium spread, s^*. In the data, the standard deviation of minimum spread is between three and five times greater under CDA than FBA, with higher values under configurations 2 and 3. The model, however, makes no statement about frequency of strategy switching in the CDA since it only demands a fixed strategy profile in aggregate. Under the FBA, everyone plays an identical strategy. Status changes are between two and 3.5 times greater under the CDA configurations, relative to FBA. The coordination difficulties that can only arise in the CDA are compatible with the CDA exhibiting a higher number of strategy changes than the FBA. The final two rows report the root mean

		Configuration 1		Configuration 2		Configuration 3	
		CDA	FBA	CDA	FBA	CDA	FBA
(a) Market conditions							
λ_I		1/3		1/5		1/2	
λ_V		1/4		1		1	
c_{speed}		0.01		0.01		0.022	
(b) Choices							
Making (%)	Experiment	54	78.1	30.2	78.8	40.1	72.9
	Equilibrium	16.7	100	16.7	100	16.7	100
Sniping (%)	Experiment	31	20.8	58.1	14.5	49.5	14
	Equilibrium	83.3	0	83.3	0	83.3	0
Speed (%)	Experiment	56.1	19.7	69	31.7	69.2	20.7
	Equilibrium	100	0	100	0	100	0
Min. Spread	Experiment	0.226	0.103	0.677	0.179	0.709	0.147
	Equilibrium	0.324	0	0.566	0	0.475	0
(c) Market stats							
$Std(P_t - P_{t-1})$	Experiment	2.51	0.561	4.62	1.00	6.68	1.11
	Equilibrium	0.241	0.623	0.276	0.281	0.235	0.281
$Std(MinSpread)$	Experiment	0.204	0.0235	0.536	0.144	0.394	0.127
	Equilibrium	0	0	0	0	0	0
Status Changes	Experiment	20.5	6.26	31.6	6.26	17.0	7.34
	Equilibrium	N/A	0	N/A	0	N/A	0
$RMSD(P_t - V_t)$	Experiment	0.347	0.212	0.512	0.410	0.460	0.381
	Equilibrium	0.223	0.141	0.329	0.237	0.372	0.312
# Transactions	Experiment	156	85.2	172	99.3	248	134
	Equilibrium	106	80	100	48	147	120
Period Profits	Experiment	.0869	.435	.603	.372	4.31	1.52
	Equilibrium	0	0	0	0	0	0

Table 1: Summary statistics for experimental data. Panel (a) Shows the parameters of the three configurations we use in the experiments. Panel (b) reports the average percentage of subjects acting as market makers, snipers, average percentage of subjects purchasing speed services, and the average minimum spread posted by market makers. Values are sampled at one-second intervals and averaged across time and subjects within each treatment. Predicted equilibrium values are reported below observed averages. Panel (c) reports summary measures related to volatility (standard deviation of price changes and minimum spread, number of strategy/role changes), informational efficiency (root mean squared deviation from price to fundamental value, RMSD), volume (number of transactions), and period profits in ECUs. For both panels, we exclude the first ten seconds of each period and the first two periods to account for starting effects.

squared deviation (RMSD) of transactions prices relative to the contemporaneous fundamental value and the average number of transactions per trading period.[2] In equilibrium, the CDA RMSD values should be close to half of the equilibrium spread and the FBA RMSD values should be close to zero.[3] In practice, the CDA values are quite close to the full equilibrium spreads and only 1.25 to 1.75 times larger than the FBA values. Finally, the number of transactions in the CDA are consistently larger than those of the FBA. In equilibrium, the CDA values should be

$$N_{trans} = 240\lambda_I + 240\lambda_V \Pr\left(J > \frac{s^*}{2}\right)\frac{5}{6}, \tag{6}$$

where 240 is the length of each trading period in seconds. The respective equilibrium values of N_{trans} for configurations 1, 2 and 3 are 106, 100 and 147, which are much lower than the observed values. Under the FBA, the number of transactions in equilibrium is $240\lambda_I$, or 80, 48 and 120, which, aside from configuration 2, are close to the observed values. This count includes transactions that occur between investors. Aside from configuration 2, those numbers are close to the observed values (Table 1). The discrepancy in C2 emerges from the high frequency of value jumps and transactions of makers sniping makers, described in Section C.

The table also shows the average net profits in ECUs, which are calculated by adding up all incurred gains and losses in each trading period and averaging those across periods and subjects (or by subtracting the endowment from the end-of-period profits). Equilibrium profits for both formats are zero. Market behavior, however, yields slightly positive profits in all formats and configurations, with configuration 3 having the highest profits for both formats. For the markets that entered this calculation (periods > 2), in only 0.5% of the 840 cases (140 markets times six traders), a trader finished the period with a negative total profit.

The BCS environment is not conducive to study allocative efficiency and therefore those measures of efficiency are less relevant here. However, those are not completely unimportant. The percentage of investors' orders that are not filled are a measure *allocative inefficiency.* Both formats predict that this measure will be equal to 100% in equilibrium. However, in the CDA, the fact equilibrium presumes coordination among traders and that is not behaviorally guaranteed implies that sometimes (when every trader is on snipe mode) some investors' orders will not be filled and lower allocative efficiency will be achieved. Indeed, there is some evidence for that behavioral conjecture. While in FBA only 0.14% of investors' orders where not filled, in the CDA this percentage is 2.57%.

It is also important to note that the purchase of speed is a measure of the deviation of Pareto efficiency. Since traders would be collectively better off by not purchasing speed at all, the more communication technologies are acquired in this context, the higher is the social waste. This is due to the fact that in this environment $V(t)$ is publicly observable and therefore faster communication technology serves no purpose in acquiring more and better information.

[2] $RMSD = \sqrt{\frac{\sum(P_t - V_t)^2}{N_{trans}}}$, where N_{trans} represents the number of transactions within a trading period.

[3] In equilibrium, FBA RMSD is slightly above zero because there are sniping-like transactions happening if, an up (down) jump occurs after $t - \delta_{slow}$ immediately followed by a buy (sell) investor.

	(1) Maker (%)	(2) Sniper (%)	(3) Speed (%)	(4) Min. Spread	(5) RMSD
Configuration 1	54.13*** (1.640)	30.89*** (2.959)	56.12*** (2.972)	0.226*** (0.0438)	0.347*** (0.00474)
Configuration 2	30.27*** (4.507)	58.11*** (3.294)	69.20*** (1.594)	0.678*** (0.114)	0.512*** (0.00918)
Configuration 3	40.03*** (2.368)	49.51*** (2.888)	69.02*** (4.498)	0.706*** (0.0374)	0.460*** (0.0170)
FBA × Config. 1	24.88*** (5.022)	-10.95* (5.507)	-36.15*** (5.399)	-0.123** (0.0439)	-0.135*** (0.0119)
FBA × Config. 2	49.08*** (6.494)	-44.23*** (6.214)	-37.82*** (3.968)	-0.502*** (0.118)	-0.102*** (0.0307)
FBA × Config. 3	33.10*** (5.255)	-35.55*** (4.309)	-48.37*** (5.024)	-0.560*** (0.0422)	-0.0791*** (0.0246)
Observations	10934	10934	10934	10934	142

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: This table reports OLS estimates for Equation (7). The dependent variables in columns 1–3 are, respectively, the average percentage of traders choosing to be makers, snipers, and subscribers to speed services. The dependent variables in columns 4 and 5 are, respectively, the average minimum spread the market-level RMSD. The regressions for all dependent variables except RMSD were conducted with data at the three-second resolution. We exclude the first ten seconds of each period and the first two periods to account for starting effects. Group-level clustered standard errors are in parentheses.

3.3 Treatment Effects

To quantify treatment effects, we estimate the following model:

$$y_{g,t} = \sum_{j=1}^{3} [\alpha_j Cj_{g,t} + \gamma_j Cj \times FBA_{g,t}] + \epsilon_{g,t}, \tag{7}$$

where $y_{g,t} \in \{Maker_{g,t}, Sniper_{g,t}, Speed_{g,t}, MinSpread_{g,t}\}$ is indexed by group and time (second in CDA and batch number in FBA), Cj is a dummy variable for market configuration $j \in \{1, 2, 3\}$, and $Cj \times FBA_{g,t}$ is the dummy variable indicating the interaction between configuration j and the FBA format. In this context, α_j captures the mean of outcome y in configuration j in the CDA format, and γ_j captures the mean FBA - CDA difference for outcome y, in configuration j. We estimate this model using a three-second time resolution, excluding the first ten seconds of each period and the first two periods of each session to account for starting effects. The estimation was done by combining data for all formats and configurations. The coefficients are OLS estimates and standard errors are cluster-robust at the group level. The regression results are reported in Table 2.

The coefficients in the second column of Table 2 show that the fraction of market makers is between 25% and 50% higher under FBA than CDA, with higher values associated with higher volatility in the asset value and fewer investors. Since market makers are constrained to limit orders of unit size, the fraction of subjects acting as market makers is a direct measure of liquidity or market depth. Thus, the results demonstrate that FBA has a positive effect on liquidity that is statistically significant. Moreover, the differences between FBA and CDA are larger in more volatile regimes: we find that the interaction coefficients, γ_2 and γ_3, satisfy $\gamma_2 > \gamma_1$ and $\gamma_3 = \gamma_1$ with $p < 0.001$.

Predatory behavior (the fraction of traders acting as snipers), reported in column 3 of Table 2, is substantially lower for all FBA configurations relative to CDA, with the difference falling in the range $[10\%, 44\%]$. Notably, the total population of snipers is much higher in the CDA for the volatile configurations (C2 and C3) , and lower in the FBA for those same configurations. Consequently, as with the fraction of market makers, larger differences are associated with higher volatility and fewer investors: again, $\gamma_2 > \gamma_1$ and $\gamma_3 = \gamma_1$ with $p < 0.001$.

Column 4 of Table 2 shows that purchases of speed services are consistently lower under FBA. Specifically, the total fraction of traders purchasing speed services under CDA falls in the range $[56\%, 69\%]$, while the FBA - CDA differences are in the range $[-48\%, -36\%]$. Equilibrium predictions, that $\alpha_j = 100$ and $\gamma_j = -100$ for $j \in \{1, 2, 3\}$, are rejected at the 1% level.

Transactions costs, measured by the average minimum spread (fifth column in Table 2) are close to the minimum increment of 0.1 ECUs in FBA, while the average CDA spreads are much higher and closer to their theoretical counterparts.[4] The FBA - CDA differences in spread were -0.123, -0.502, and -0.560 ECUs, respectively for C1, C2 and C3, and serve as upper bounds for the differences, since they exclude the many investor-to-investor transactions at the fundamental value (zero spread) that occur in the FBA when an equal number of buy and sell investors arrive within one batch. In general, these differences closely track equilibrium values: for C2 we do not reject the null hypothesis that $\alpha_1 = .566$ and $\gamma_1 = -0.46$, while for C1 and C3, the minimum spread values are close to, but statistically distinct from, equilibrium levels. Despite this, a key comparative static across CDA configurations holds in the data: CDA-C2 and CDA-C3 both have statistically higher minimum spread relative to CDA-C1.

The results of these regression estimations are robust to changes in the time resolution of the analysis. In particular, the same specifications with data at three-second time resolution (not reported) yield nearly identical results in coefficients and standard errors to those in table 2.

The last column of Table 2 displays a regression of the root mean squared deviation following this specification:

$$RMSD_h = \sum_{j=1}^{3} [\alpha_j Cj_h + \gamma_j Cj_h \times FBA_h] + \epsilon_h. \tag{8}$$

This regression is fit at the level of trading period and group (h), therefore it utilizes 142 observations.[5] Consistent with predictions, RMSD is lower under FBA. Specifically, RMSD under CDA is in the range $[0.347, 0.512]$, while the FBA - CDA differences are in the range $[-0.135, -0.0791]$.

In sum, the comparative statics of the differences between the two formats predicted by the model are confirmed in the data: Relative to the CDA, the FBA has more makers, fewer snipers, fewer traders purchasing speed technology, lower minimum spreads and lower RMSDs. For all relevant coefficients, except in one, the differences between formats are statistically significant at the 1% level.

[4] Our lab interface implemented a minimum spread of 0.1 ECUs.

[5] Six treatments with four groups each, trading in periods 3-8, minus two group-periods because of the session that stopped after period seven.

3.4 Transitory Market Dynamics

As in real financial markets, subjects in our experiments observed information regarding the trading environment in real time. This included explicit information on the total number of market makers and their (unattributed) spreads, as well as implicit information on the presence of snipers, the volatility of the asset, and the frequency of investor arrivals. To understand the dynamics of the market and the possible effects of transitory changes in the environment on subjects' decisions, we fit a vector autoregression of the form:

$$\boldsymbol{y}_t = \boldsymbol{a} + \boldsymbol{\Phi} \boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{9}$$

$$\boldsymbol{y}'_t = [\%Sniper_t, \%Speed_t, MinSpread_t, Turbulence_t], \tag{10}$$

where, as before, $\%Sniper_t$ is the average fraction of subjects choosing to be snipers during time interval t, $\%Speed_t$ is the average fraction of players choosing to purchase speed services during time interval t, and $MinSpread_t$ is the average minimum spread in the market over time interval t. As the new magnitude under analysis, $Turbulence_t$ is defined as the ratio of number of recent price changes to the number of recent investor arrivals, $\frac{N_{V,t}}{N_{I,t}}$, during time interval t [6]. Since N_V and N_I are exogenously determined by calibrated Poisson processes, we constrain the VAR so that the last element of $\boldsymbol{a}$ and the last row of $\boldsymbol{\Phi}$ are equal to zero and so that the last element of $\boldsymbol{\varepsilon}_t$ is equal to $Turbulence_t$. We set the time interval of the regression to be 3 seconds, as this is the natural interval of the FBA treatment (the length of the batch), and is reasonable interval over which to measure transitory effects in the CDA [7].

From the perspective of a market maker, turbulence measures the countervailing forces of price volatility (increased sniping costs) with rate of investor arrivals (increased market making income). Further, the theoretical counterpart λ_V/λ_I plays a direct role in the determination of the BCS equilibrium for the CDA, as seen in Equation (3). Equation (5) shows that only λ_V has an impact on the FBA equilibrium. The constrained VAR in Equations (9) and (10) not only allows us to measure the transitory effects of players' choices in one period on subsequent behavior, it also captures the transitory effects of turbulence on each outcome variable, net of related effects on other choices.

Table 3 reports parameter estimates of the constrained VAR(1) for both CDA and FBA. The first row of panel (b) shows that the fraction of snipers, the fraction of subjects purchasing speed and the turbulence all have a positive and significant (at least at the 5% level) relationship with subsequent decisions to snipe. The minimum spread, on the other hand, has a negative relationship, at the 5% level. Interestingly, the only lagged variable to be significantly related to speed purchases, is the fraction of agents purchasing speed, whereas minimum spread is positively and significantly (again, at least at the 5% level) related to all variables except speed purchases (no significance). Panel (b), however shows that the only statistically significant relationships under FBA are variables lagged with themselves. Specifically, turbulence has no impact on subsequent behavior in the FBA. Altogether, the results show that very-short term, innovations in market conditions impact behavior in the CDA. Such effect of transient market changes do not exist in the FBA.

Figure 4 depicts impulse responses for the estimated VARs reported in Table 3 for a horizon of

[6] We do not include the fraction of market makers in the VAR since it is highly co-linear with the fraction of snipers.

[7] We also evaluated the VAR for longer time intervals, with diminishing effects in the interval length, suggesting that subjects' reactions to the changing environment are short term. This is corroborated in the impulse responses that we report below.

	(a) CDA				
	Constant	$\%Sniper_{t-1}$	$\%Speed_{t-1}$	$MinSpread_{t-1}$	$Turbulence_{t-1}$
$\%Sniper_t$	10.1***	0.720***	0.0651**	-3.14**	0.141**
	(2.15)	(0.0354)	(0.0320)	(1.24)	(0.0670)
$\%Speed_t$	10.3***	0.0480	0.813***	-0.494	-0.00220
	(1.90)	(0.0313)	(0.0283)	(1.09)	(0.0593)
$MinSpread_t$	-0.00389	0.00250**	0.000877	0.656***	0.00665***
	(0.0611)	(0.00101)	(0.000911)	(0.0352)	(0.00191)
	(b) FBA				
	Constant	$\%Sniper_{t-1}$	$\%Speed_{t-1}$	$MinSpread_{t-1}$	$Turbulence_{t-1}$
$\%Sniper_t$	4.66***	0.752***	0.0347	-10.1**	-0.0442
	(1.03)	(0.0317)	(0.0331)	(4.13)	(0.0529)
$\%Speed_t$	5.73***	-0.0129	0.752***	2.98	0.0182
	(0.995)	(0.0305)	(0.0319)	(3.97)	(0.0509)
$MinSpread_t$	0.0683***	-0.000182	0.0000620	0.529***	-0.000132
	(0.0102)	(0.000313)	(0.000327)	(0.0408)	(0.000523)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Parameter estimates for the constraint VAR(1) in Equations (9) and (10). Standard errors are reported in parentheses. Panel (a) reports CDA estimates and panel (b) reports FBA estimates.

20 three-second periods, or 1 minute of clock time. In each case, the impulse responses measure the effect of a one-unit increase in the impulse variable, holding all other variables constant. The responses in both figures are congruent with the parameter estimates reported in Table 3. Specifically, under FBA shocks to each of the variables only have autocorrelative effects and no cross-correlative effects in subsequent time periods. Under CDA (panel (a) of Figure 4), (1) a transitory 1% increase in the fraction of snipers leads to a small (no more than 0.003 ECUs), positive and significant increase in the minimum spread for about 40 seconds, (2) a transient 1% increase in the fraction of subjects purchasing speed services results in an increase in the fraction of snipers by as much as 0.1% for no more than 10 seconds, (3) a transitory increase in the minimum spread by 1 ECU leads to a significant decline in the fraction of snipers by as much as 4% for nearly 30 seconds, and (4) a unit increase in turbulence results in a very short (less than 5 seconds), but significant increase in the fraction of snipers and a longer (a little over 20 seconds) and significant increase in the minimum spread, by as much as 0.006 ECUs. When we run the same analysis only using periods 5-8 (not reported), we find essentially the same results. This leads us to suggest that the discussed short-term dynamics is a permanent feature of each formats and not caused by different starting effects or learning patterns. As a whole, these VAR results show that relative to the CDA, the FBA attenuates behavioral responses to short-term shocks in market conditions.

4 Conclusions

We use laboratory experiments to empirically study and compare the performance of the CDA and FBA. The environment for our experiments follows the model of Budish et al. [2015], where a single asset is traded on a single exchange and two exogenous processes generate incentives to trade: changes in a publicly-observed fundamental value of an asset and the arrival of market orders from noise traders (*investors*). In our experiment, human participants (acting as traders) tune algorithms that trade on

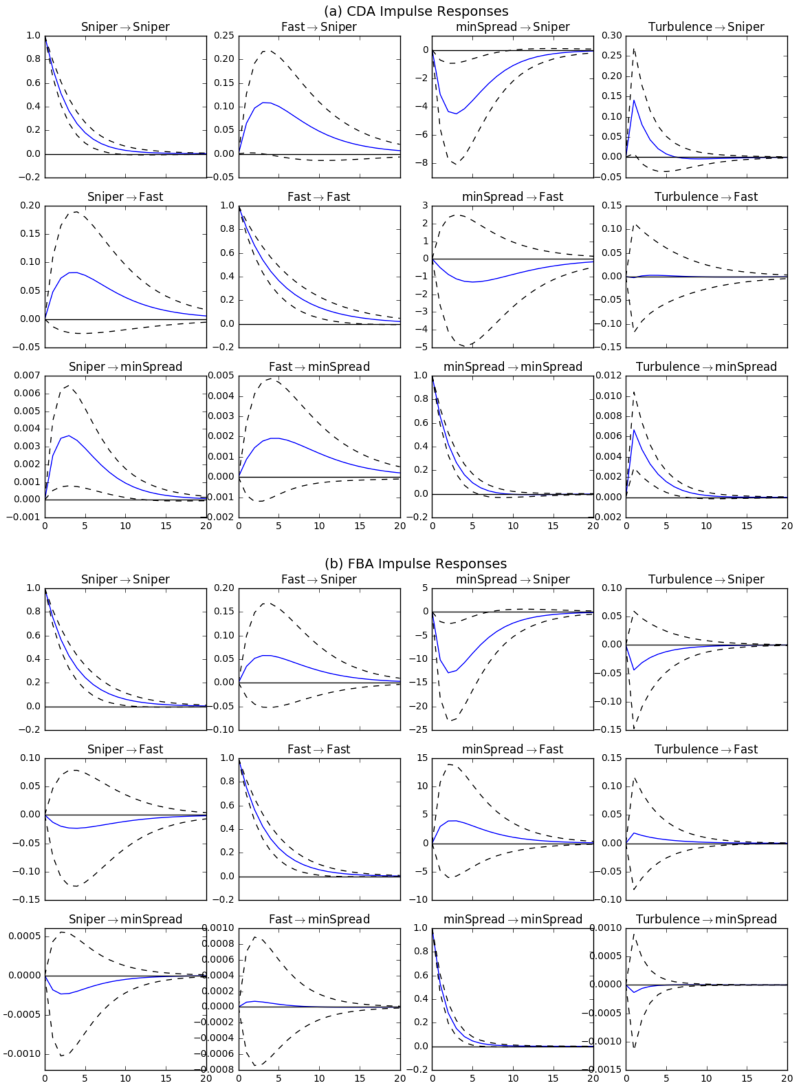


Figure 4: Unit impulse responses for the estimated VAR under CDA (panel a) and FBA (panel b).

their behalf. For a cost, traders have access to a technology that reduces latency of messaging with the (remote) exchange. We emulate modern financial markets by developing an electronic architecture in which information and trading occur at millisecond time granularity and which follows the Nasdaq OUCH messaging protocol. Each of the market formats (CDA and FBA) is studied under three market conditions, varying the degree of volatility in the market and the cost of technology.

We find that, compared to the CDA, the FBA exhibits higher levels of liquidity, less predatory behavior, less investment in communication technology, lower transactions costs, and higher informational efficiency. Specifically, there are between 25-50% more market markers under the FBA than the CDA, 10-44% fewer aggressive traders seeking to trade on stale information, and 36-49% fewer traders purchasing speed technology. Further, the average minimum spread among market makers is substantially lower in the FBA, volatility of transactions prices and spreads is lower, and deviations in transactions prices from the underlying asset value are lower as well. As an additional measure of interest, we estimate the sensitivity of traders' behavior to transitory shocks in several market environment variables and find that the sensitivity is statistically absent in the FBA, despite the fact that such shocks have statistically significant short-run effects on behavior in the CDA.

We also find that market behavior is closer to BCS predictions in more turbulent markets (e.g., C2 relative to C1). There are two possible explanations for that finding. First, it could be that indeed the issues raised in Budish et al. [2015] become stronger in stressed markets. Second, alternatively, it could be that markets with higher turbulence provide more opportunities for learning given that information arrives more frequently. Although our evidence suggest that learning cannot explain the finding, our experiment is not designed to discriminate clearly between those two explanations. We leave the study of this important aspect to future research.

Taken together, our results suggest that the frequent batch auction is a welfare improving mechanism for allocating trade within a financial market. Such market designs are of interest to policy makers and regulators, as they reduce both volatility and dead-weight loss, and shift lost surplus to investors with fundamental portfolio needs. Further, our results suggest that exchange operators may find such alternative market designs promising, as they will inherently attract liquidity from fundamental investors and encourage fast traders to forgo expenditure on fast communication technology and instead focus their attention on liquidity provision.

We recognize that our results are as much a test of the BCS model as a comparison of the FBA and CDA formats, and that they are only relevant to the extent that the model broadly characterizes actual strategies and interactions by agents in the market (which we think it does). Beyond the relatively simple strategy space for agents in this environment, it would be useful to test the robustness of the FBA mechanism to more sophisticated behavior and complex preferences. Such behavior might include traders that place multiple-unit orders (which either deepen liquidity or have substantial price impact), investors that intelligently shred orders through time, and reactive algorithms (for both makers and snipers) that respond to other traders' strategies, rather than to changes in the environment alone.

In going from the BCS environment to more complex an realistic environments, our experimental approach will generate insights for both further theory development and the implementation of field experiments. Along these lines, we believe at least two paths of research will be fruitful in advancing our understanding of financial market design: first, studying alternative market formats (e.g, Kyle and Lee [2017] and Aldrich and Friedman [2017]) in comparable laboratory environments, and second, conducting controlled field experiments that exhibit more realistic and complex features. Indeed, within the framework of a larger research project, we intend to implement a public experiment (in the form of a tournament) in which we will study different financial market institutions in a less stylized environment.

We also leave for future research the theoretical and empirical analysis of the protection that different market formats provide to unsophisticated players. The coordination involved in the equilibrium of the CDA format in the BCS model and the higher complexity of the best response function relative to

References

Eric M. Aldrich and Daniel Friedman. Order Protection through Delayed Messaging. *Working Paper*, 2017.

Eric M. Aldrich, Joseph A. Grundfest, and Gregory Laughlin. The Flash Crash: A New Deconstruction. *Working Paper*, 2016.

Dan Ariely, Axel Ockenfels, and Alvin E Roth. An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics*, pages 890–907, 2005.

Markus Baldauf and Joshua Mollner. Trading in Fragmented Markets. *Working Paper*, 2015a.

Markus Baldauf and Joshua Mollner. High-Frequency Trading and Market Performance. *Working Paper*, 2015b.

N Bershova and D Rakhlin. High-Frequency Trading and Long-Term Investors: A View from the Buy Side. *Working Paper*, 2012. ISSN 1556-5068. doi: 10.2139/ssrn.2066884. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:High-Frequency+Trading+and+Long-Term+Investors+:+A+View+from+the+Buy+Side#0.

Bruno Biais, Thierry Foucault, and Sophie Moinas. Equilibrium Fast Trading. *Working Paper*, 2014.

Johannes Breckenfelder. Competition Between High-Frequency Traders, and Market Quality. *Working Paper*, 2013.

Jonathan Brogaard and Corey Garriott. High-Frequency Trading Competition. *Working Paper*, 2015.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High Frequency Trading and Price Discovery. *Review of Financial Studies*, 27(8):2267–2306, 2014.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. Price Discovery without Trading: Evidence from Limit Orders. *Working Paper*, pages 1–51, 2015.

Eric Budish, Peter Cramton, and John Shim. Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye. *American Economic Review: Papers & Proceedings*, 104(5):418–424, 2014. ISSN 00028282. doi: 10.1257/aer.104.5.418.

Eric Budish, Peter Cramton, and John Shim. The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *The Quarterly Journal of Economics*, 130(November):1547–1621, 2015. doi: 10.1093/qje/qjv027.Advance.

T. N. Cason and D. Friedman. Price formation and exchange in thin markets: A laboratory comparison of institutions. In Peter Howitt, Elisabetta de Antoni, and Axel Leigonhufvud, editors, *Money, Markets and Method: Essays in Honour of Robert W. Clower*, number February, chapter 8, pages 155–179. Elgar, 1999. URL http://www.krannert.purdue.edu/faculty/cason/papers/fourinst.pdf.

Timothy N. Cason and Daniel Friedman. Price formation in double auction markets. *Journal of Economic Dynamics and Control*, 20:1307–1337, 1996.

Timothy N. Cason and Daniel Friedman. Price Formation in Single Call Markets. *Econometrica*, 65 (2):311–345, 1997.

Timothy N. Cason and Daniel Friedman. A Comparison of Market Institutions. In Charles R. Plott and Vernon L. Smith, editors, *Handbook of Experimental Economics Results*, volume 1, chapter 33, pages 264–272. North-Holland, Amsterdam, 2008. ISBN 9780444826428. doi: 10.1016/S1574-0722(07) 00033-9.

Songzi Du and Haoxiang Zhu. What is the Optimal Trading Frequency in Financial Markets? *Review of Economic Studies*, 2017.

Thierry Foucault. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*, 2:99–134, 1999.

Daniel Fricke and Austin Gerig. Too Fast or Too Slow? Determining the Optimal Speed of Financial Markets *. *Working Paper*, 2015. URL http://ssrn.com/abstract=2363114.

Daniel Friedman. Privileged Traders and Asset Market Efficiency: A Laboratory Study. *The Journal of Financial and Quantitative Analysis*, 28(4):515–534, 1993.

Daniel Friedman. laboratory financial markets. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, second edition, 2008.

Ben Greiner. Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125, 2015.

Björn Hagströmer and Lars Nordén. The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770, November 2013. ISSN 13864181. doi: 10.1016/j.finmar.2013.05.009. URL http://linkinghub.elsevier.com/retrieve/pii/S1386418113000256.

Joel Hasbrouck and Gideon Saar. Low-latency trading. *Journal of Financial Markets*, 16(4):646–679, November 2013. ISSN 13864181. doi: 10.1016/j.finmar.2013.05.003. URL http://linkinghub.elsevier.com/retrieve/pii/S1386418113000165.

Terrence Hendershott and Ryan Riordan. Algorithmic trading and the market for liquidity. *Journal of Financial and Quantitative Analysis*, 48(4):1001–1024, 2013. ISSN 0022-1090. doi: 10.1017/S0022109013000471. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2001912.

Nicholas Hirschey. Do High Frequency Traders Anticipate Buying and Selling Pressure? *Working Paper*, 2013. ISSN 1556-5068. doi: 10.2139/ssrn.2238516. URL http://papers.ssrn.com/sol3/Delivery.cfm?abstractid=2238516.

Charles A. Holt. Industrial Organization: A Survey of Laboratory Research. In John H. Kagel and Alvin E. Roth, editors, *The Handbook of Experimental Economics*, volume 1, chapter 5, pages 349–444. Princeton University Press, Princeton, 1995.

Boyan Jovanovic and Albert J Menkveld. Middlemen in Limit Order Markets. *Working Paper*, 2015.

Albert S Kyle and Jeongmin Lee. Toward a fully continuous exchange. *Oxford Review of Economic Policy*, 33(4):650–675, 2017. doi: 10.1093/oxrep/grx042. URL +http://dx.doi.org/10.1093/oxrep/grx042.

Katya Malinova, Andreas Park, and Ryan Riordan. Do retail traders suffer from high frequency traders? *Working Paper*, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2183806. URL http://qed.econ.queensu.ca/pub/faculty/milne/322/IIROC_FeeChange_submission_KM_AP3.pdf.

Albert J Menkveld and Marius a Zoican. Need for Speed? Exchange Latency and Market Quality. *Working Paper*, pages 1–51, 2015.

Charles N. Noussair and Steven Tucker. Experimental Research on Asset Pricing. *Journal of Economic Surveys*, 27(3):554–569, 2013. doi: 10.1111/joes.12019.

Alvin E Roth and Axel Ockenfels. Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet. *American Economic Review*, 92 (4):1093–1103, 2002.

Alvin E. Roth and Xiaolin Xing. Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions. *The American Economic Review*, 84(4):992–1044, 1994.

Alvin E. Roth and Xiaolin Xing. Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists. *Journal of Political Economy*, 105(2):284–329, 1997.

Shyam Sunder. Experimental Asset Markets: A Survey. In John H. Kagel and Alvin E. Roth, editors, *The Handbook of Experimental Economics*, chapter 6, pages 445–500. Princeton University Press, Princeton, 1995.

Sarah Zhang and Ryan Riordan. Technology and Market Quality: The Case of High Frequency Trading. *ECIS 2011 Proceedings*, (Paper 95), 2011.

Appendices

A Calibration

Aldrich and Friedman [2017] obtain proprietary data from the IEX exchange for the month of December, 2016. IEX classifies each participant as either an "agency" or "proprietary" trader, the former being the class of traders with a fundamental interest to buy or sell assets (i.e. to maintain an inventory for portfolio reasons). Aldrich and Friedman [2017] report that IEX agency transactions comprised 10,498,518 shares of the S&P 500 exchange traded fund (ticker SPY) during the 21 trading days or $21 \times 6.5 \times 60 = 8190$ trading minutes during December, 2016. Since the median trade size is the minimum block of 100 shares, this amounts to $10,498,518/100 \approx 105,000$ total trades during the month, or $105,000/8190 = 12.82$ investor arrivals per minute, or roughly 1 investor arrival every 4.68 seconds. Although IEX represents only a small fraction of equities market share, we believe that most fundamental traders (investors) will utilize IEX in conjunction with other equities exchanges, and hence that their arrival rates would be suggestive of aggregate investor arrival intensities. The result is that in raw time (we discuss time rescaling below), the Poisson intensity parameter for investor arrivals is $\tilde{\lambda}_I = 1/4.68$.

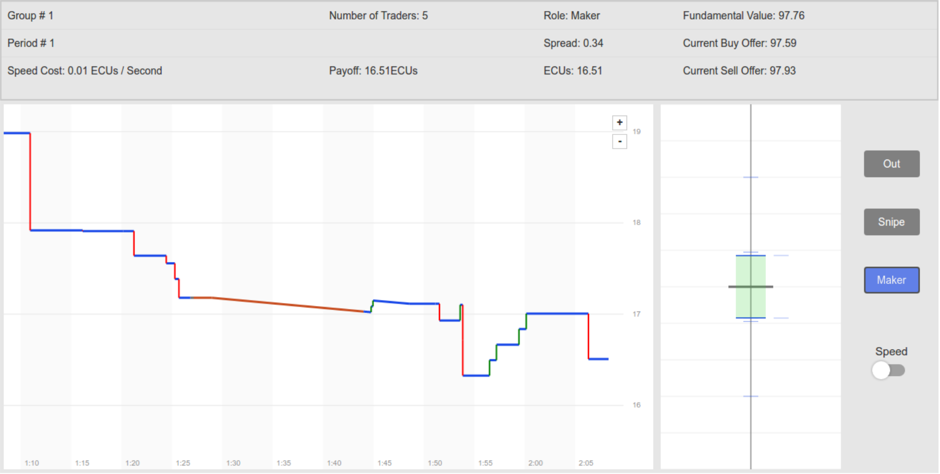
To calibrate $\tilde{\lambda}_V$, the intensity of the Poisson process governing jumps in the fundamental value, we utilize SPY quotation data at Nasdaq, which, given its liquidity and overall market share, is a good surrogate for the SPY national best bid and offer (NBBO). Our sample covers the period 16 June – 11 September, 2014. There are 26,216,524 quotations in the 62-day period, which comprises 1,450,800,000 milliseconds during trading hours, or approximately 1 quote every 55 milliseconds. Defining a jump as any midpoint price change of magnitude at least \$0.01 over the period of four quotations, or 220 milliseconds, resulted in a median of approximately 3978 jumps per day, or one jump every 5.88 seconds (assuming 23,400 seconds during the 6.5 hour equities market trading day). Hence, $\tilde{\lambda}_J = 1/5.88$, prior to time rescaling.

Following Aldrich et al. [2016], we assume the trade-time distribution of asset price changes, $\Delta V(t)$, is Gaussian with mean zero. Using the SPY data above, we find that $Std(\Delta V(t)) = \$0.007$, or slightly less than the minimum spread of \$0.01. However, given the preponderance of liquidity at SPY best bid and offer and the fact that the minimum spread is set by the SEC, the unconstrained equilibrium spread is widely considered to be less than \$0.01. This suggests that the standard deviation of value changes should be of similar magnitude to the equilibrium minimum spread. As the magnitude of the scale parameter is otherwise arbitrary, we set $\Delta V(t) \sim \mathcal{N}(0, \sigma = 0.5)$, resulting in a scale parameter that is of the same approximate magnitude as the CDA equilibrium spreads. Further, the choice region for maker spreads in the experiment was set to encompass a region of 4σ around the fundamental value.

We set $\lambda_V = 1/4$ in our baseline calibration, which leads us to interpret 1 second of lab time as $\lambda_V/\tilde{\lambda}_V = 5.88/4 = 1.47$ seconds of raw financial market time. Consequently, we interpret a single four-minute experimental period as approximately $4 \times 1.47 = 5.88$ minutes of financial market time, and the full eight-period session as approximately $8 \times 4 \times 1.47 \approx 47$ minutes of market time. Further, rescaling the investor intensity parameter to experimental time results in $\lambda_I = 1.47/4.68 \approx 1/3$, which is the value of our baseline calibration.

To calibrate the cost of fast communication technology, c_s, we use the pricing schedule of McKay Brothers LLC, a premier microwave transmission service. To transmit a single symbol on the longhaul route between the CME data center in Aurora, IL to an equities data center in New Jersey costs \$10,600 per month, or about \$0.02 per second (assuming 22 trading days per month, and 6.5 trading hours per day). Scaling to experimental time, this results in about \$0.015 per second. In our baseline calibration we set $c_s = \$0.01$ per second.

B FBA User Interface



Group # 1	Number of Traders: 5	Role: Maker	Fundamental Value: 97.76
Period # 1		Spread: 0.34	Current Buy Offer: 97.59
Speed Cost: 0.01 ECUs / Second	Payoff: 16.51ECUs	ECUs: 16.51	Current Sell Offer: 97.93

Figure 5: FBA user interface in BCS environment.

C Off-Equilibrium Sniping

In the BCS equilibrium under CDA, sniping transactions occur between one of the $N - 1$ snipers and the single maker. In practice, if there is more than one maker (off-equilibrium) at the time of a jump in the fundamental value, the maker whose orders are repriced first may snipe the maker who is repriced after. These maker-to-maker sniping transactions are not explicitly dealt with in Budish et al. [2015] and raise the question as to whether they could alter the equilibrium. Empirically, we find the prevalence of these transactions is substantial. However, augmenting the model with this strategy does not extend the set of equilibria, which remains unique up to the aggregate composition of making and sniping strategies.

We illustrate this type event with an example. Consider a market operating under the CDA. Suppose at $t = 0$ the value of the asset is $V = 100$ and only two traders are present in the market as makers. Let us refer to these traders as M1 and M2, and assume that they both have a spread of $s = 2$ (bids at 99 and offers at 101), and that M1 has purchased speed services (operating at a latency of 100 ms) and M2 has not (operating at a latency of 500 ms). If the value jumps to $V = 105$ at $t = 700$ ms, M1's algorithm will submit messages to the exchange, updating her bid and offer to 104 and 106, respectively. Those messages will be received at the messaging server at $t = 800$ ms and immediately passed to the order book, where M1's new bid of 104 will cross with M2's stale offer of 101. Under typical exchange rules, the order is filled at 101, resulting in a profit (loss) of $105 - 101 = 4$ to M1 (M2).[8]

[8] Similar events occur less frequently when makers (fast or slow) are filled by an investor and the asset value changes

In the equilibrium under FBA, on the other hand, sniping transactions occur between investors and any of the N makers when the value jump occurs too late for makers to update quotes before batch end. In practice, if some makers purchase speed technology (off-equilibrium), maker-to-maker sniping can occur. However, since these events can only happen when jumps occur close to the end of the batch, their prevalence is much less common than in the CDA. Additionally, if one or more traders decide to act as snipers (off-equilibrium), both slow and fast makers will be sniped by investors, fast makers and fast snipers, in that order. As shown below, this is the pattern we find in the data.

Table 4 reports frequency counts of snipes under each of the CDA and FBA configurations. In each case, the number of snipes is decomposed by the roles of the traders participating in the sniping transaction. While the parties of each transaction are clearly defined in the CDA, the same is not true of the FBA: there is no specific attribution of which traders transact with each other in a call-type auction. To make such attributions, we paired sniped traders (those with negative profits) with non-sniped traders in the following priority: (1) fast snipers, (2) fast makers, and (3) investors. Since slow makers and slow snipers never have an opportunity to snipe in the FBA, we make no such attributions. Although our particular attribution order may be somewhat arbitrary, it corresponds to the correct attribution in batches with a single transaction and the aggregate counts are comparable to those of the CDA.

	Investor		Fast Maker		Slow Maker		Fast Sniper		Slow Sniper		Total	
	CDA	FBA	CDA	FBA	CDA	FBA	CDA	FBA	CDA	FBA	CDA	FBA
(a) Configuration 1												
Fast Maker	32	14	641	10	7	NA	432	0	8	NA	1,120	24
Slow Maker	22	193	870	112	28	NA	413	55	51	NA	1,384	360
(b) Configuration 2												
Fast Maker	38	49	397	61	25	NA	1,885	0	21	NA	2,366	110
Slow Maker	29	271	536	639	39	NA	1,473	181	83	NA	2,160	1092
(c) Configuration 3												
Fast Maker	135	77	388	31	18	NA	2,102	2	23	NA	2,666	110
Slow Maker	69	448	416	479	22	NA	1,048	158	153	NA	1,708	1085

Table 4: Snipes decomposition. Frequency counts of snipes under each of the CDA and FBA configurations, decomposed by the role of the sniper and the trader being sniped.

The values in Table 4 demonstrate that sniping is much more prevalent under the CDA, with the bulk of CDA snipes done by fast snipers or fast makers. Interestingly, consistent with the off-equilibrium insight at the beginning of this section, fast makers account for a larger share of sniping in CDA configuration 1, relative to CDA configurations 2 and 3. This latter result is due to the higher population of snipers under the latter two configurations (see Figure 3 and Table 1). Under FBA, the number of times that fast makers are sniped is about 20 times less than under CDA, and slow makers are sniped roughly 2 to 4 times less. Further, we make the bulk of FBA attributions to fast makers and investors rather than fast snipers.

As successful sniping of any form occurs infrequently under FBA, the consequences of unintentional maker snipes are more important under CDA. That is, makers have access to similar profit opportunities as snipers, with the additional benefit (cost) of investor transactions (being sniped). We detail the

multiple times during the interval in which new orders are being routed to the exchange. We do not focus on these, as they account for very little total volume.

trade-offs to this expanded strategy space below, and prove that the model equilibrium is unchanged. Despite the identical equilibrium, it is possible that the perceived benefit of maker-to-maker snipes may partly explain why our experimental observations are attenuated relative to predicted BCS equilibrium values.

C.1 Equilibrium with Maker-Snipers in the CDA

We now show that extending the BCS model to explicitly account for maker-to-maker sniping leaves the CDA equilibrium intact.

Proposition C.1. *Consider an augmented BCS model that allows repriced maker orders to transact with stale limit orders posted by other traders. The equilibrium of the augmented model is identical to that of Budish et al. [2015].*

Proof We sketch the proof for the case of the CDA with endogenous entry. Suppose an equilibrium exists with N trading firms acting as market makers, quoting the same spread, s, and all purchasing fast communication technology. The profit to each firm would be

$$\lambda_I \cdot \frac{s}{2N} - \lambda_V \cdot \Pr\left(J > s\right) \cdot \mathbb{E}\left[J - \frac{s}{2} | J > s\right] \cdot \Pr(\mathit{Sniped})$$
$$+\lambda_V \cdot \Pr\left(J > s\right) \cdot \mathbb{E}\left[J - \frac{s}{2} | J > s\right] \cdot \Pr(\mathit{Sniping}) = c_{speed} \quad (11)$$
$$\Rightarrow \lambda_I \cdot \frac{s}{2N} = c_{speed}. \quad (12)$$

The LHS of Equation (11) has positive and negative terms related to sniping which cancel, resulting in Equation (12). We examine the terms on the LHS of Equation (11) and compare them with the typical profit condition for a maker in the BCS equilibrium (Equation 1):

1. Investor profits under the all-maker equilibrium are shared among all N firms: $\lambda_I \cdot \frac{s}{2N}$. In the BCS equilibrium, a single trading firm acts as market maker and captures all of the profit alone: $\lambda_I \cdot \frac{s}{2}$.

2. Makers can snipe other makers, but it requires a larger jump to do so: the jump must be at least s in magnitude in order for the bid (offer) of a sniping maker to cross with the stale offer (bid) of another maker. In the BCS equilibrium, a jump of magnitude $s/2$ suffices for a sniper to transact with a stale maker quote. In both cases, the sniping profit $J - \frac{s}{2}$. Thus, the second term on the LHS of Equation (11) breaks down in the following way: (i) λ_V is the jump intensity, (ii) $\Pr(J > s)$ is the probability that a jump is big enough for the maker to be sniped, (iii) $\mathbb{E}\left[J - s | J > \frac{s}{2}\right]$ is the expected sniping profit (conditional on sufficiently large jump), and (iv) $\Pr(\mathit{Sniped})$ is the probability that a maker is sniped. The same terms in Equation (1) are: (i) λ_V, (ii) $\Pr\left(J > \frac{s}{2}\right)$, which is larger than $\Pr(J > s)$, (iii) $\mathbb{E}\left[J - \frac{s}{2} | J > \frac{s}{2}\right]$, which is less than $\mathbb{E}\left[J - \frac{s}{2} | J > s\right]$, and (iv) $\frac{N-1}{N}$, which is an upper bound for $\Pr(\mathit{Sniped})$ (see below).

Letting $\Pr(\mathit{Sniping})$ denote the probability of a maker successfully sniping another maker, it can be readily shown that the $\Pr(\mathit{Sniping}) = \Pr(\mathit{Sniped})$ (it is a version of the High School Prom Theorem). Furthermore, these probabilities are lay withing the open interval $\left(\frac{int(N/2)}{N}, \frac{N-1}{N}\right)$. Thus, the losses attributed to sniping in Equation (11) are exactly offset by the gains, resulting in a cancellation of the second and third terms on the LHS of Equation (11).

We now explore the profitability of two possible deviations: (1) a single maker quoting a more narrow spread and (2) a single maker switching roles to act as a pure sniper. Maintaining the notation s^* for the BCS equilibrium spread, we let $\hat{s}$ denote the all-maker equilibrium spread. For the first deviation (narrower spread) to be profitable, the following would need to hold

$$
\begin{aligned}
\lambda_I \cdot \frac{\hat{s}}{2N} < \lambda_I \cdot \frac{\hat{s}-\varepsilon}{2} & \\
- \lambda_V \cdot \Pr\left(J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}-\varepsilon}{2} | J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right] \cdot \frac{N-1}{N} & \qquad (13) \\
+ \lambda_V \cdot \Pr\left(J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}-\varepsilon}{2} | J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right] \cdot \Pr(\mathit{Snipe}) & \qquad (14) \\
\Rightarrow N(\hat{s}-\varepsilon) - \hat{s} > 2 \cdot \frac{\lambda_V}{\lambda_I} \cdot \Pr\left(J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}-\varepsilon}{2} | J > \frac{\hat{s}+(\hat{s}-\varepsilon)}{2}\right] \cdot (N-1). & \qquad (15)
\end{aligned}
$$

Intuitively, the deviating maker would capture all investor profits at the expense of increased probability, $\frac{N-1}{N} > \Pr(\mathit{Sniped})$, of getting sniped. In Equation (15), we have eliminated a positive constant which does not affect the inequality. Taking the limit as $\varepsilon \to 0$, Equation (15) can be expressed as

$$
\hat{s} > 2 \cdot \frac{\lambda_V}{\lambda_I} \cdot \Pr\left(J > \hat{s}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}}{2} | J > \hat{s}\right]. \qquad (16)
$$

Alternatively, if a single maker deviates to act as a sniper, the resulting profit would be

$$
\lambda_V \cdot \Pr\left(J > \frac{\hat{s}}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}}{2} | J > \frac{\hat{s}}{2}\right] \cdot \frac{N-1}{N} = c_{speed} \qquad (17)
$$

Thus, a sniper deviation is profitable if

$$
\begin{aligned}
\lambda_V \cdot \Pr\left(J > \frac{\hat{s}}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}}{2} | J > \frac{\hat{s}}{2}\right] \cdot \frac{N-1}{N} > \lambda_I \cdot \frac{\hat{s}}{2N} & \qquad (18) \\
\Rightarrow \hat{s} < 2 \cdot \frac{\lambda_V}{\lambda_I} \cdot \Pr\left(J > \frac{\hat{s}}{2}\right) \cdot \mathbb{E}\left[J - \frac{\hat{s}}{2} | J > \frac{\hat{s}}{2}\right] \cdot (N-1). & \qquad (19)
\end{aligned}
$$

To determine if a profitable deviation exists, we compare Equations (16) and (19). If at least one of the conditions is always satisfied, the all-maker equilibrium cannot be supported. This is the case if the RHS of Equation (19) is always greater than the RHS of Equation (16):

$$
\begin{aligned}
\Pr\left(J > \hat{s}\right) \cdot \mathbb{E}\left[J - \tfrac{\hat{s}}{2} | J > \hat{s}\right] \quad &< \quad \Pr\left(J > \tfrac{\hat{s}}{2}\right) \cdot \mathbb{E}\left[J - \tfrac{\hat{s}}{2} | J > \tfrac{\hat{s}}{2}\right] \cdot (N-1) \\
\Rightarrow \Pr(J > \hat{s}) \tfrac{\int_{\hat{s}} (z - \frac{s}{2}) 2\phi(z) dz}{\Pr(J > \hat{s})} \quad &< \quad \Pr(J > \tfrac{\hat{s}}{2}) \tfrac{\int_{\frac{\hat{s}}{2}} (z - \frac{s}{2}) 2\phi(z) dz}{\Pr(J > \frac{\hat{s}}{2})} \cdot (N-1) \\
\Rightarrow \textstyle\int_{\hat{s}} (z - \frac{s}{2}) 2\phi(z) dz \quad &< \quad \textstyle\int_{\frac{\hat{s}}{2}} (z - \frac{s}{2}) 2\phi(z) dz \cdot (N-1) \\
\Rightarrow \textstyle\int_{\hat{s}} (z - \frac{s}{2}) 2\phi(z) dz \quad &< \quad \textstyle\int_{\frac{\hat{s}}{2}} (z - \frac{s}{2}) (N-1) 2\phi(z) dz \qquad (20) \\
\Rightarrow \textstyle\int_{\hat{s}} (z - \frac{s}{2}) 2\phi(z) dz \quad &< \quad \textstyle\int_{\frac{\hat{s}}{2}}^{\hat{s}} (z - \frac{s}{2}) (N-1) 2\phi(z) dz + \int_{\hat{s}} (z - \frac{s}{2}) (N-1) 2\phi(z) dz \\
\Rightarrow 0 \quad &< \quad \textstyle\int_{\frac{\hat{s}}{2}}^{\hat{s}} (z - \frac{s}{2}) (N-1) 2\phi(z) dz + \int_{\hat{s}} (z - \frac{s}{2}) (N-2) 2\phi(z) dz.
\end{aligned}
$$

Since the LHS of Equation (20) is always positive for $N \geq 3$, we conclude that the all-maker equilibrium cannot be supported. Further deviations can be derived inductively, resulting in the equilibrium as stated in Budish et al. [2015].

D Equilibrium Market Statistics

D.1 $Std(P_t - P_{t-1})$

The variance of price changes is defined to be

$$Var\left(P_t - P_{t-1}\right) = E\left[\left(P_t - P_{t-1}\right)^2\right] - E\left[P_t - P_{t-1}\right]^2 .. \tag{21}$$

Under CDA,

$$\begin{aligned} E\left[\left(P_t - P_{t-1}\right)^2\right] &= \lambda_I\left(\frac{1}{2}0^2 + \frac{1}{2}s^2\right) + \lambda_J Pr\left(J > \frac{s}{2}\right)\left(\frac{1}{2}0^2 + \frac{1}{2}s^2\right) && (22) \\ &= \frac{s^2}{2}\left(\lambda_I + \lambda_J Pr\left(J > \frac{s}{2}\right)\right) && (23) \end{aligned}$$

and

$$E\left[P_t - P_{t-1}\right]^2 = \frac{s^2}{4}\left(\lambda_I + \lambda_J Pr\left(J > \frac{s}{2}\right)\right)^2 . \tag{24}$$

The terms on the right-hand-side of Equation (**??**) are a result of the fact that all price differences are either zero (bid to bid or offer to offer) or a full spread (bid to offer or offer to bid) independent of whether they are attributed to investors or snipes (at the time of jumps). Substituting these terms into Equation (21) implies

$$Std(P_{t+1} - P_t) = \sqrt{Var\left(P_t - P_{t-1}\right)} = s\sqrt{\left(\frac{A}{2} - \frac{A^2}{4}\right)} \tag{25}$$

and,

$$Std\left[P_t - P_{t-1}\right] = s\sqrt{\frac{A}{2} - \frac{A^2}{4}} \tag{26}$$

where: $A = \lambda_I + \lambda_J Pr(J > \frac{s}{2})$.

Under FBA, absent jumps, $P_t = V_t$ since all traders act as makers and quote zero spreads. This implies that price changes are perfectly correlated with value changes and that the standard deviation of $P_t - P_{t-1}$ should be identical to the standard deviation of $V_t - V_{t-1}$, but scaled by the probability that a value jump occurs in the batch and at least one investor arrives:

Regarding RMSD, we approximate it as:

$$RMSD = \sqrt{E\left[(P_t - V_t)^2\right]} \tag{27}$$

$$\begin{aligned} E\left[(P_t - V_t)^2\right] &= \lambda_I\left(\frac{s}{2}\right)^2 + \lambda_J Pr[J > \frac{s}{2}]E[(J - \frac{s}{2})^2 | J > \frac{s}{2}] \\ &= \lambda_I\left(\frac{s}{2}\right)^2 + \lambda_J \int_{\frac{s}{2}} (J - \frac{s}{2})^2 dF \end{aligned} \tag{28}$$

$$RMSD = \sqrt{\lambda_I \left(\frac{s}{2}\right)^2 + \lambda_J \int_{\frac{s}{2}} (J - \frac{s}{2})^2 dF} \tag{29}$$

Equilibrium Market Stats for the FBA

Here t refers to the auction occurring at clock time τt. Following the formula for the variance of a compound Poisson process, we have:

$$Std\left[P_t - P_{t-1}\right] = \sigma\sqrt{\tau\lambda_V \left(1 - e^{-\tau\lambda_I}\right)} \tag{30}$$

For the RMSD, since prices are different from $V(t)$ only when there is investor imbalance and the V jump occurs after δ_{slow} seconds before then end of the auction, we have that the probability of P being different than V is given by $\delta_{slow}\lambda_V \cdot \left(1 - e^{-\tau\lambda_I}\right)$. Also, conditional to the jump event happening at r, and considering that the price is either $V(r^-)$ or $V(r)$, then the magnitude $E[(P_t - V_t)^2]$ is simply $E[(V(r) - V(r^-))^2]$ or σ^2.

$$E\left[(P_t - V_t)^2\right] = \delta_{slow}\lambda_V \left(1 - e^{-\tau\lambda_I}\right) \cdot \sigma^2 \tag{31}$$

Which means:

$$RMSD = \sigma\sqrt{\delta_{slow}\lambda_V \left(1 - e^{-\tau\lambda_I}\right)} \tag{32}$$