ETF PRIMARY MARKET STRUCTURE AND ITS EFFICIENCY

Ewelina Anna Zurowska

A DISSERTATION

in

Finance

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Nikolai Roussanov, Moise Y. Safra Professor of Finance

Graduate Group Chairperson

Nancy R. Zhang, Ge Li and Ning Zhao Professor of Statistics and Data Science

Dissertation Committee

Nikolai Roussanov, Moise Y. Safra Professor of Finance
Donald B. Keim, John B. Neff Professor of Finance
David Musto, Ronald O. Perelman Professor in Finance

ACKNOWLEDGEMENT

I am deeply indebted to my advisers Donald Keim, David Musto, and especially Nick Roussanov, for their guidance and continuous support.

ABSTRACT

ETF PRIMARY MARKET STRUCTURE AND ITS EFFICIENCY

Ewelina Anna Zurowska

Nikolai Roussanov

The primary market of many US registered ETFs exhibits an oligopolistic structure, which is shown to have relevant implications for the pricing efficiency of those financial products. I show that the entry of an additional Authorized Participant (AP) corresponds to a decrease in the magnitude of ETF price deviations from Net Asset Value (NAV) of at least one basis point in ETFs with high primary market concentration. I build a dynamic equilibrium model of ETF primary market arbitrage that describes the trade-off faced by monopolistically competitive APs between waiting for mispricing to widen and pre-empting competitors from eliminating it. In the model, the creation unit size is shown to be an important friction driving the entry decision and, therefore, the magnitude of mispricing. Indeed, in the data, around one-third of all primary market transactions amount to one creation unit, suggesting that it is often a binding constraint. ETF split events and the creation unit size changes help to identify shocks to the dollar value of creation unit size empirically. I show that by cutting the creation unit size in half, mispricing decreases by almost two basis points, a magnitude consistent with that implied by my quantitative model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

The ETF sector has been snowballing over the last two decades. While still smaller than mutual funds, the total assets under management for the US ETF listings reached more than $5 trillion at the end of 2020. As the popularity of ETFs has recently grown so rapidly among retail and institutional investors, some believe that ETFs could potentially overtake the traditional open-end mutual fund sphere in the future. One of the major differences is that ETF shares are publicly traded, providing investors with instant (i.e., intraday) liquidity, as well as serving as a convenient tool for hedging and speculative trading. In contrast to open-end funds, the market price of ETF shares is typically close but does not coincide with the net asset value (NAV). The magnitude of mispricing (either discount or premium to NAV) is typically substantially smaller than that of closed-end fund shares;[1] nonetheless, the high ETF trading volume of $25 trillion can potentially translate into annual mispricing of ETF transactions of up to $40 billion.[2] The mispricing also tends to widen substantially during periods of market stress.[3]

In this paper, I study the role of the ETF primary market structure in driving the deviations of ETF prices from NAV. The primary market is a key feature of ETF design that is meant to ensure that ETF share prices do not deviate far from NAV. Specifically, ETFs allow

[1] For example, average absolute mispricing for Blackrock MuniAssets Closed-End Fund (ticker MUA) during Jan 2017—Oct 2021 was 4.7%, versus 0.32% for SPDR Nuveen Bloomberg High Yield Municipal Bond ETF (ticker HYMB). Both funds invest in high-yield federal tax-exempt US municipal bonds. Also, Engle and Sarkar (2006) suggest that the premiums and discounts of international closed-end funds are larger and more persistent than those observed for international ETFs. Even stronger example of the importance of the primary market with well designed built-in arbitrage for pricing efficiency of structured products presents the recently very popular Grayscale Bitcoin Trust (GBTC), US registered grantor trust that holds only bitcoin. It features a closed-end-like structure as its shares are traded in the OTC market with very high daily volumes. Yet, due to the lack of redemptions and occasional halts on creations, the GBTC shares notoriously traded at the highly persistent premiums or discounts of around 15% on average in 2019-2021 according to https://www.theblockcrypto.com/data/crypto-markets/structured-products/premium-of-gbtc.

[2] Estimate number computed as daily ETF volume multiplied by the end-of-day absolute mispricing, 2019—20 annual average.

[3] For example, during the flash crashes as documented in Borkovec et al. (2010), Ben-David et al. (2017), Kay (2009), or Madhavan (2016), or during the March 2020 market crash as documented in Haddad et al. (2020) or Todorov (2021).

designated entities called authorized participants (APs) to swap the underlying securities for ETF shares in big blocks called creation units, typically of a size of 50,000 ETF shares. This mechanism enables ETF size adjustment in response to investor demand, but also, crucially, it allows the APs to profitably arbitrage away differences between an ETF's share price and its underlying NAV. I analyze to what extent the imperfect competition between the APs in the primary market of the ETF limits the effectiveness of the arbitrage mechanism. In addition, I investigate the role that the creation unit size, which is exogenous to the APs arbitrage decisions and for a long time was constrained by regulation, plays in shaping the competitiveness of the ETF primary market and, consequently, the magnitude of mispricing.

First, using a new dataset created by parsing annual N-CEN reports from EDGAR, a new regulatory filing for ETFs, I analyze the actual primary market structure of the ETFs registered in the United States. I show that a large portion of the US-listed ETFs exhibits an oligopolistic structure in the number of active authorized participants. I also show that the limited competition structure of the ETF primary market, as measured by the market concentration index or the number of active authorized participants, contributes to larger deviations of an ETF's price from its NAV, known as a mispricing. The biggest concern in this exercise is the endogeneity problem. The competition argument suggests that the more authorized participants there are in a particular ETF market, the more efficient this market should be and the smaller the observed ETF mispricings. However, the wider the discounts and premiums a particular ETF shows on the secondary market, the more attractive the primary market is for the arbitrage incentives, and potentially the higher the number of active authorized participants we can observe. We can see that those two effects are mitigating each other. Thus, the endogeneity issue, as such, will bias the competition effect towards zero, leaving the estimate still relevant as the lower bound on the magnitude of the competition effect.

Second, I show that high creation unit size could impede the efficient pricing of ETFs. Creation units are blocks of ETF shares that can be created or redeemed by the AP. An

AP must create or redeem shares in multiples of those blocks. Partial units are not allowed. This size was originally mandated to be 50,000, and even though it's no longer required, most ETFs still use the same value. Because 30% of all redemptions and creations are for the minimum block size of one, this raises the question of whether the creation unit is too large. ETF issuers rarely change the creation unit size, but it does happen on occasion, however mostly accompanying splits. Because not all splits are accompanied by creation unit size change, I can also use these events to analyze their impact on ETF mispricings. The results confirm that the decreases in the effective level of the creation unit size improved the ETF primary market efficiency and significantly lowered ETF mispricings.

Lastly, in the theoretical part of my analysis, I propose a model of the ETF primary market that helps to analyze how the competition in this market affects the evolution of mispricing and what effect the creation unit policy may have on it. In particular, I model the intertemporal choice of the AP to either immediately exercise the arbitrage opportunities arising from mispricing on the ETF market or wait until the gap widens to generate even higher arbitrage profits. I also show that effective competition may discourage APs from waiting, encouraging them to create or redeem at the lower levels of mispricing, thus contributing to the improvement of ETF market pricing efficiency. Although the model does not address the question of the optimal level of the creation unit, as this would require broader discussion also including the perspective of the ETF issuer, it may however guide ETF issuers in their decision-making by providing an insightful analysis of a potential primary market response to different creation unit size policies.

Studying the mispricing of ETFs is important for several reasons. First, the importance of ETF mispricing manifests itself not only through the high level of ETF trading volumes but also through the actions of the Securities and Exchange Commission (SEC). To protect individual investors and educate them about the risks that ETFs pose in terms of purchasing and selling ETF shares at a price different than NAV, the SEC in the new ETF Rule from December 2019[4] imposes a variety of disclosures on ETF websites concerning the level of mispricing. In particular, ETFs are obliged to disclose the previous-day premium and discount, how frequently premiums and discounts were present during the last calendar year and calendar quarter, and even a line graph showing the historical evolution of the ETF's share premium or discount over the last year. And in the event that a premium or a discount exceeds 2%, ETFs must provide a statement explaining the contributing factors.

Second, mispricings tend to increase significantly during times of stress, posing an additional level of risk for an investor. The most striking evidence may be the flash crash of 2010, when 70% of canceled trades were ETF trades, or the flash crash of August 2015, when extraordinary levels of mispricings were observed, often larger than 20%, even for ETFs tracking very popular indexes, such as, iShares Core S&P 500 (IVV). As mentioned in Hu and Morley (2018), the ETF market failure on those days showed that the ETF arbitrage mechanism is more fragile than anyone had thought and that additional assumptions are necessary for such mechanisms to be effective. The authors also cites an extract from a BlackRock publication, which referred to a letter written to the SEC expressing similar concerns over the efficiency of the ETF arbitrage mechanism. Additionally, the events during the 2020 pandemic and the Fed's novel ETF buying program gave a new role to ETFs in monetary policy. Historically large ETF price deviations from NAV observed in April 2020, especially in the bonds market, and documented in Haddad et al. (2020), highlight the need to better understand the ETF primary market structure and its efficiency. In good times, the ETF arbitraging infrastructure of the primary market seems to work fairly well, with mispricings rarely exceeding 2%. However, when the APs step back from trading, for example, because they are exposed to more risk than they are comfortable with, which is likely to happen in times of stress, the disconnect between an ETF's price and the price of the securities that it is supposed to mimic can quickly manifest itself. As pointed out in Madhavan and Sobczyk (2015), the staleness of NAV due to the way it is computed could

[4]Final Rule on ETF adopted by the SEC in 2019 under the Investment Company Act of 1940 simplifying the ETF regulatory process, available at: https://www.sec.gov/rules/final/2019/33-10695.pdf

be one of the reasons behind the large mispricings observed in times of stress. However, if the primary market were more efficient, it could potentially weaken the staleness effect of NAVs and thus lessen the observed mispricings in times of stress.

The relevance of studying the competition structure in the ETF primary market may not seem obvious at first. Some argue that the ETF primary market does not suffer from limited competition among APs, because when we count all the financial institutions that sign the authorized participant agreement with the ETF sponsor, this number is often quite large. In earlier studies, such as Antoniewicz and Heinrichs (2015) or Madhavan (2016), it was reported to be 36.[5] However, it is relevant that signing an authorized participant agreement gives an AP the right, but not the obligation, to participate in the primary market. Given rather limited barriers to signing up as an AP, what we observe in the data is that although many APs sign up with the ETF issuers, only few are transacting in their primary markets, putting the ETF's built-in arbitrage to use. Thus, another useful statistic of the ETF primary market is the number of APs that are actually transacting in the primary market. Using N-CEN filing records, we can categorize an AP as active for a particular ETF if it created or redeemed ETF shares at least once during the reporting period. The average number of active APs is about four for all ETFs, suggesting an oligopolistic market structure, at least in a subset of the ETFs. Generally, the ETFs with more assets under management and the ETFs with domestic underlying securities have more APs that are active than those with smaller assets under management (AUM) and illiquid or foreign underlying securities.

There can be a variety of reasons why an AP may choose not to transact in the individual ETF primary market on a regular basis: they may not consider it their primary line of business or main profit-generating strategy, they may have skills and specializations in other parts of the market, they may not have enough balance sheet inventory or capacity, they may

[5] The author believes that this number could be overstated in earlier studies that relied on surveys filled out by ETF managers. Prior to the introduction of the N-CEN filing requirement, the list of authorized participants was not well maintained by ETFs, often showing authorized participants who had registered a long time ago and might not even exist anymore or whose names were entered multiple times because of a name or ownership structure change. In this paper, the author shows that the average number of registered APs is 23 as per N-CEN filings.

have capital constraints, or they may have registered only to secure themselves the facility to create or redeem ETF shares in case there is such a request from their client (whom they would charge an intermediary fee). As such, it is hard to believe that a non-active agent whom we did not see transacting in the primary market for 12 months is following this market closely, just waiting to jump back in as an arbitrageur. Indeed, even if there might be a few, we can still observe the oligopolistic structure in this market. Also, as many of the APs are large financial institutions, often operating in a variety of businesses, it is not unlikely that they are occasionally incapable of engaging in the ETF creation and redemption due to financial constraints, or simply not willing to engage due to better investment opportunities. Constantly engaging in the process of creating/redeeming shares to correct for the demand shocks in the secondary market may not be the best use of their capital.

Some suggest that ETF mispricing is held tight also thanks to arbitrageurs in the ETF secondary market, thus limiting the importance of the competition structure in the primary market. Although there may indeed be many investors trading on the secondary market to extract profits from their trading strategies centered around ETFs and their mispricings, their trading strategies are not truly arbitrage and rely heavily on the assumption that there is an effective ETF primary market in the first place, with APs that will ensure that the ETF prices converge to NAV, closing the mispricings sooner or later. And the less efficient the primary market of the ETF is, the longer it may take to close the mispricing gaps.

The APs themselves have noticed potential issues with the competitiveness level of the ETF primary markets. In the comment written by Jane Street Capital, which specializes in ETF markets, to the SEC from 2018,[6] we can read that the authorized participants serve as a vital intermediary between the liquidity providers and the fund itself and that *"the ETF's arbitrage mechanism will function best when the arbitrage is competitive, low-cost and open to all liquidity providers."* An especially worrying situation might occur when a single authorized participant may also act as a liquidity provider and complete the arbitrage

[6] JaneStreetCapital (2018), https://www.sec.gov/comments/s7-15-18/s71518-4467045-175801.pdf

opportunity while denying other liquidity providers the possibility of creating or redeeming at a reasonable price.

There are three major features of the primary market that can potentially create a friction for the arbitrage and limit the competitiveness level of the primary market: creation fees charged by the ETF issuer, the large size of the creation unit, and the composition of the creation or redemption basket. First, the creation fees appear to be of negligible value, as suggested by the market practitioners and also shown empirically later in this paper. Second, creation unit size has been highlighted by market practitioners as being quite relevant, especially for funds holding illiquid or hard-to-access constituents. Also, for a very long time, there was quite a high floor of 50,000 shares or $1 million on its level, and only recently have ETFs been allowed to lower the creation unit size if they perceive it as beneficial. Third, the basket composition has also been highlighted by market practitioners as a relevant element, since the basket composition determines how easy it is to collect the underlying creation basket or dispose of the underlying redemption basket, how deep into an order book one needs to go when constructing such baskets, and how much of a price impact it entails. The natural question is thus, can relaxing the above-mentioned frictions help improve the ETF pricing efficiency as measured by the price deviations from NAV? In this paper, I address this question in relation to the first and second frictions. The third friction is left for future research.

The majority of the research so far has focused on the relationship between ETFs and the underlying securities, for example, how the ETF market affects the volatility, liquidity, or co-movement of the returns of the underlying constituents. Although each of those analyses, whether theoretical or empirical, relies on the arbitrage mechanism that is built into the ETF market model, there has so far been little investigation into this arbitrage mechanism's efficiency and whether it relies on a particular structure of the primary market. To the best of my knowledge, neither the competition structure among the authorized participants in the primary market of ETFs nor the creation unit size, and their importance for the ETF pricing dynamics, has received much attention in the recent literature.[7] In this paper, I thus contribute to the literature by investigating these aspects of the primary market structure of ETFs.

This paper contributes to the classic canon on limits to arbitrage by showing that the dual-tier structure of ETFs, together with the intertemporal choice of an AP, poor competition in the primary market, as well as high levels of primary market transaction units may pose limits to arbitrage other than those suggested in other works by Shleifer and Vishny (1997), Long et al. (1990), Harrison and Kreps (1978), Gromb and Vayanos (2010), as well as more recent and ETF specific work by Pan and Zeng (2019), which shows that not only the liquidity of underlying assets but also the AP's inventory capacity may affect the ETF arbitrage mechanism.

This paper also extends the literature on ETFs and the risks that they pose. In particular, to the paper by Ben-David et al. (2018), which shows that there is a volatility transformation, Evans et al. (2021), which investigates the link between operational shorting and fail-to-deliver events in the ETF primary market; and that of, Brown et al. (2020), which suggests that primary market transactions provide signals of non-fundamental demand shocks. It also contributes to the papers on the persistence of ETF mispricing, such as work by Madhavan and Sobczyk (2015), Lettau and Madhavan (2018) and Madhavan (2016), and to the literature modeling the ETF market structure, e.g. Malamud (2016). Additionally, it contributes to the novel literature branch on the creation and redemption baskets in the ETF sphere, along with the recent paper by Shim and Todorov (2021) documenting the strategic choices made by ETF managers when constructing the creation and redemption baskets. Other papers worth mentioning in the ETF sphere are An et al. (2021), who study the competition structure among the ETF index providers; Brogaard et al. (2021), who study the relevance of choosing the index components; and Petajisto (2017), who studies the levels of mispricing

[7] However, there is a very recent paper by E. Gorbatikov and T. Sikorskaya titled *"Two APs Are Better Than One: ETF Mispricing and Primary Market Participation"* analyzing the network features of AP—ETF links.

among a group of similar ETFs.

Lastly, this paper also contributes to the literature on the competition among arbitrageurs on the more broadly defined markets, such as papers by Kondor (2009), Attari et al. (2005), Attari and Mello (2006), Fardeau (2021), Gromb and Vayanos (2018), or studying a specific markets as in case of a paper by Kozhan and Tham (2012) on FX markets or paper by Mercadal (2021) on electricity trades.

CHAPTER 2

Institutional Framework

An exchange-traded fund is similar to a closed-end fund (CEF) because it can be traded like a stock throughout the day. However, a CEF has a fixed number of shares, making it more susceptible to demand shocks. Conversely, an ETF does not have a fixed number of shares. In this sense, it is more similar to an open-end fund (OEF), whose shares are created by the mutual fund company once a day to meet investors' demand. ETFs achieve this by using authorized participants, typically bigger financial institutions, with exclusive rights from the issuer to control the share supply via redemption and creation so that the market price stays close to the underlying NAV.

Each day, the ETF issuer is required to publish the creation and redemption baskets for the next trading day, which contain the specific list of names and quantities of securities and cash. Usually, the baskets would represent the ETF's portfolio holdings at a pro-rata slice. Occasionally, the basket may be composed of a representative sample, cash, or other securities.[8] When an AP wants to create shares, it submits such an order to the ETF agent, the entity that manages the fund. The ETF issuer has to approve the order and confirm it with the AP. The cut-off time for placing such orders is 4 pm for the vast majority of ETFs; however, for some bond or foreign ETFs, the cut-off can be earlier in the day. The AP can create or redeem shares only in big blocks called creation units, with a typical creation unit size of 50,000 shares. Upon the delivery of the specified creation basket to the ETF, the newly created ETF shares are transferred to the AP. Similarly, upon the delivery of the ETF shares worth the creation unit size, the securities specified in the redemption baskets are released to the APs. This type of creation or redemption process is described as in-kind and, to a large extent, is the reason for the important tax benefits that ETFs offer to a fund and its investors and to transacting agents.[9]

[8] Especially in cases of ETFs with thousands of holdings or illiquid assets (bonds or foreign securities), the creation and redemption basket may differ from the portfolio holdings constituents.

[9] The paper by Antoniewicz and Heinrichs (2014) provides a detailed description of the creation and

The ETF regulatory framework that is relevant to this paper comprises the rules that used to surround the size of the creation unit. With the original 2001 Class Letter,[10] the Securities and Exchange Commission (SEC) prescribed that *"Fund Shares are only to be issued or redeemed in Creation Unit aggregations of 50,000 shares or more. The value of each Creation Unit must be at least $1 million at the time of issuance."* This prescription was later modified in the 2006 Class Letter[11] to *50,000 shares or such other amount where the value of a Creation Unit is at least $1 million at the time of issuance.* Thus, unsurprisingly, for decades the vast majority of ETFs had their creation unit size set to 50,000.

However, an order issued in December 2017[12] eliminated the creation unit size requirement for equity index-based ETFs. About a year later, with a no-action letter from September 2018,[13] the SEC effectively further eliminated the minimum creation unit size requirement for Fixed-Income and Combination ETFs.

---

redemption process, including all clearing and settlement details.

[10]Exemptive Relief for Exchange Traded Index Funds from August 17th, 2001, available at https://www.sec.gov/divisions/marketreg/mr-noaction/etifclassrelief081701-msr.pdf

[11]Class Relief for Exchange Traded Index Funds from October 24th, 2006, available at https://www.sec.gov/divisions/marketreg/mr-noaction/etifclassrelief102406-msr.pdf

[12]Order Granting Limited Exemptions from Exchange Act Rule 10b-17 and Rules 101 and 102 of Regulation M to Certain Index-Based ETFs Pursuant to Exchange Act Rule 10b-17(b)(2) and Rules 101(d) and 102(e) of Regulation M from December 7th, 2017, available at https://www.sec.gov/rules/exorders/2017/34-82234.pdf

[13]No-action letter response from September 20th, 2018, available at https://www.sec.gov/divisions/marketreg/mr-noaction/2018/invesco-092018-regm-101-102.pdf

CHAPTER 3

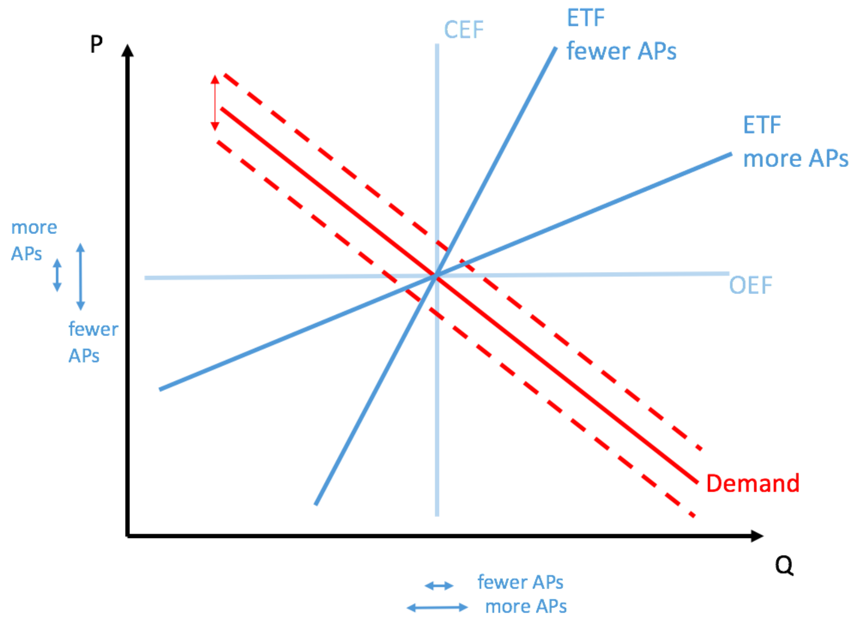
PQ model of ETF primary market

As mentioned earlier, ETFs can be considered as a hybrid between closed-end funds (CEFs) and open-end funds (OEFs). The simple PQ model can help compare these three types of investment products to each other and shed some light on the importance of primary market efficiency. Figure 3.1 presents an illustration of the simple PQ model. Q denotes the number of shares in the market. P denotes the mispricing (price relative to NAV).

In the case of OEFs, the supply curve can be viewed as perfectly elastic, as the number of OEF shares can fluctuate freely — any investor can directly purchase or redeem any number of shares from the fund. As the shares can always be purchased or redeemed at NAV, the mispricing is theoretically zero (for simplicity, I omit possible redemption fees). Thus, as the demand for an OEF shifts, only the number of shares adjusts, and the price stays equal to NAV. In the case of CEFs, the supply curve is perfectly inelastic, as the number of shares is fixed at the fund's creation, and generally, new shares cannot be created, nor old shares redeemed. Investors can only acquire or sell CEF shares on the secondary market by finding a prospective buyer or seller. Thus, in the light of demand shocks to CEF shares in the secondary market, the price can deviate far from the NAV, and large mispricing levels are not uncommon for CEF shares.

The ETF share supply is somewhat elastic, however not as much as that of OEFs. It can only be adjusted by the activity of APs, and only in big chunks, called creation units. However, the majority of investors need to head to the secondary market to purchase or sell ETF shares, thereby setting their price. Thus, due to liquidity shocks (or demand shocks) in the secondary market, an ETF's price can deviate from the fundamental value (NAV), causing it to trade at a premium or a discount. The more APs are active in the ETF primary market, the more elastic the ETF supply curve will be, and thus, the smaller the mispricing fluctuations should be and the bigger the fund flows should be. In contrast, the fewer APs

Figure 3.1: PQ model of ETFs



Source: Author's own elaboration.

are active in the ETF primary market, the less elastic the ETF supply curve will be, and thus, the smaller the mispricing fluctuations should be and the bigger the shares fluctuation should be. In the long term, the price of an ETF should be equal to its fundamental value; thus, mispricing should be equal to zero.

CHAPTER 4

Empirical Analysis

4.1. Data

Data on ETFs were sourced from various providers. First, I use the SEC's EDGAR system to parse the data on ETF authorized participants from N-CEN filings. Form N-CEN is a new filing that the registered investment companies have to report on an annual basis since June 2018. ETF issuers are required to file Form N-CEN by 75 days after their fiscal year-end. Form N-CEN, which replaced the N-SAR filing, collects information on a fund's financial and legal structure, and more importantly, it requires ETF issuers to provide detailed information on their primary market structure, such as the list of authorized participants, their aggregate activity in the reporting period, the composition of creation/redemptions orders between cash and in-kind, and the fees charged for primary market transactions. Parsing the filings makes it possible to create the following variables of interest: number of registered APs and number of active APs.[14] I also compute the proxy for primary market concentration for each ETF using the data provided in the N-CEN filing. Since the ETF issuer has not only to list the APs by name but also to provide the aggregate value of shares created and redeemed by each individual AP, I calculate the Herfindahl—Hirschman index (HHI) for each ETF as the sum of the squared market shares of each AP.

$$HHI_i = \sum_j \left(\frac{S_i^j}{\sum_j S_i^j} \right)^2$$

where S_i^j denotes the j-th AP's aggregate transactions with the i-th ETF. Under this specification, HHI can take values between 0 and 1, with 1 corresponding to the case with only one active AP. When the number of active APs is 0, the above ratio is undefined; however, in the empirical analysis, I assume it is equivalent to the monopoly by setting HHI to 1.

[14] An active AP is defined as an AP that performed at least one creation or redemption transaction with the ETF issuer in the reporting period.

Since the N-CEN filing requirement went into effect rather recently and ETFs had to provide detailed information on their primary market activity for the first time, initial reports were often not filled correctly, and a large number of amended reports were submitted. Thus, in my database, I keep the most recent N-CEN filing for the specified reporting period if an amendment was submitted instead of the initial N-CEN filing. The constructed database contains all filings submitted between July 2018 and August 2021. During this period, 2,507 ETFs filed at least one N-CEN filing, 2,031 filed at least two consecutive filings, 1,631 filed N-CEN at least three times, and 22 filed N-CEN at least four times. The latter group was for the ETFs that switched the month of their fiscal year-end, resulting in some N-CEN filings being filed for a reporting period shorter than 12 months. To remove any biases that could stem from unequal reporting periods, when computing the market structure measures in my empirical analysis, I restrict my data to only the reports covering a full 12-month period, giving me a total of 6,142 year-long N-CEN reports covering 1,645 ETFs with three year-long reports, 356 ETFs with two year-long reports, and 494 ETFs with one year-long report. This allows me to explore not only the cross-sectional variation in the primary market structure among ETFs but also the time variation.

Additionally, I collect daily data on prices, volumes, bid—ask spreads, shares adjustment factors for splits from CRSP, and fund flows from Bloomberg, and historical creation unit size, assets under management, net asset values, asset class of ETF, and portfolio holdings from ETFGlobal. I also use ETFGlobal to identify the related index for passive ETFs in the fund pair-matching empirical exercise. As some of the data sources overlaps, I combine them to limit the number of missing values. I construct my final dataset by merging daily data on ETFs with SEC aggregate information on their primary market, and I keep only those daily observations that overlap with the N-CEN reporting period for each of the ETFs individually.[15]

[15]Please see Table A.1 in the appendix for more details.

Table 4.1: ETF primary market concentration across N-CEN filings

This table presents the measures of ETF primary market concentration. Number of registered APs represents the number of all APs listed by ETF issuer in N-CEN filing. Each individual AP entry made by an ETF is counted as a seperate AP, even if it falls under the same ultimate holding company. Number of active APs represents the number of APs who created or redemeed shares at least once during the reporting period. HHI concentration index was computed for each ETF's primary market as a sum of squares of ratios of the individual AP's creation (redemption or primary market transactions volume) transaction relative to the total level of creations (redemptions or primary market transactions volume) for that particular ETF. The observation unit is N-CEN filing. Data span N-CEN filings filed July 2018 — August 2021. Reports that span the period of less than 12 months are excluded. In case number of active APs is zero, the HHI index is set to 1.

	count	mean	sd	min	p5	p10	p25	p50	p75	p90	p95	max
Number of registered APs	6,142	23	14	1	2	4	12	23	37	44	44	45
Number of active APs	6,142	4.4	3.3	0	1	1	2	4	5	8	11	24
Frequency of active APs	6,142	.32	.32	0	.038	.056	.11	.18	.35	1	1	1
HHI for creations	6,142	.58	.28	.093	.21	.26	.35	.51	.85	1	1	1
HHI for redemptions	6,142	.64	.3	.088	.21	.27	.37	.56	1	1	1	1
HHI for volume	6,142	.52	.26	.092	.2	.24	.32	.45	.66	1	1	1
Creation Unit Size ['000]	6,141	57	44	5	.25	25	50	50	50	100	100	600

Source: Author's computations.

4.2. Empirical Results

To better understand the primary market structure of ETFs, I computed the basic summary statistics using the N-CEN filings that are presented in Table 4.1. The number of registered APs varies between 1 to 45 with the mean at 23. The number of active APs[16] is much smaller and varies between 0 and 24, with an average value of 4. We can also see that on average, only every fifth registered AP engages in creation or redemption activity, as measured by the median. The HHI computed for the total volume of transactions in the primary market of an ETF is on average half, which is equivalent to a duopoly market structure with two players with identical market share. Given that the average number of active APs is larger than two, we could also interpret this by saying that, on average, an ETF has one prominent AP and three less active ones. Table 4.1 also reports the distribution of creation unit size across filings. The majority of ETFs have a creation unit size of 50,000 shares.

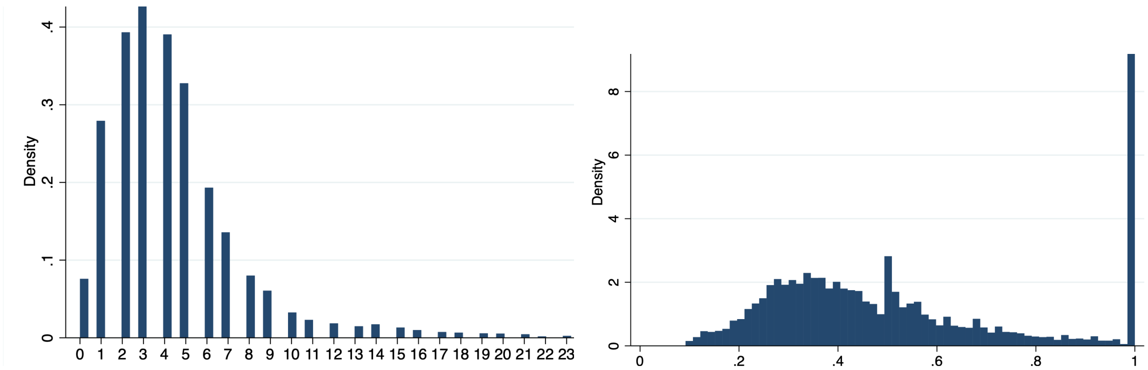
The median value of a portion of the basket swapped in-kind for creation orders is 97%, and for the redemption orders, it is 98%. In Table A.3, I show more detailed statistics on the ETFs' creation and redemption basket composition across filings. Generally, there is a tendency for the creation or redemption baskets of ETFs with less liquid underlying securities, for example, fixed-income ETFs, to have higher cash components, and smaller and newly created ETFs also tend to accept creation baskets with higher cash components to grow more quickly.

Figures 4.1a and 4.1b depict histograms of the number of active APs and the volume HHI. We can easily see that only few ETFs would have more than 10 active APs and the most common number of APs per ETF is 3. Also, for at least 10% of ETFs, the primary market has no more than 1 active AP and for at least 25% of ETFs, no more than 2 APs are active. This suggests that a large number of ETFs may suffer from the consequences of an oligopolistic structure in this market.

[16] An active AP is defined as an AP that created or redeemed ETF shares at least once during the reporting period of N-CEN filing with a particular ETF.

Figure 4.1: Distributions of ETF market concentration

The graphs below shows the distribution for the number of active APs and the HHI market concentration using ETF primary market transaction volumes across ETFs. Data spans the N-CEN filings submitted between July 2018 - August 2021. Reports that cover the period shorter than a year were excluded.



(a) Histogram of active APs (b) Histogram of HHI volume

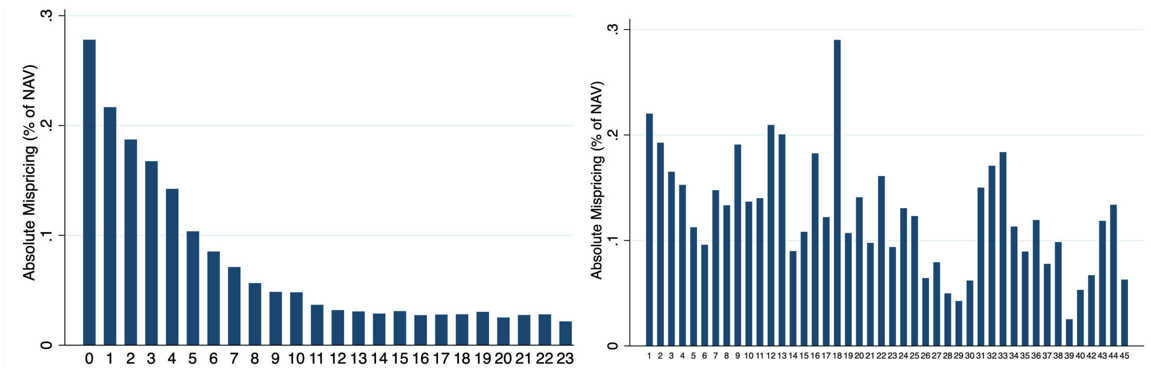
Source: Author's computations.

Figures 4.2a and 4.2b show the level of observed ETF mispricings across the number of registered and active APs. We can see a very strong and smooth decreasing relationship with the number of active APs that seems to vanish when the number of active APs crosses into double digits, and a much choppier and barely visible relationship with the number of registered APs. This suggests that the number of active APs seems to better capture the nature of the concentration structure in ETF primary market.

The juxtaposition of the creation fees against the potential arbitrage profits, as expressed by the level of absolute relative mispricing, is presented in Table 4.2. Here, we can see that the creation fees charged by the ETF issuers for the primary market transactions are indeed marginal, with a median value of 0.35 basis points and a mean value of 0.7 basis points if related to one creation unit. At the same time, the level of the absolute mispricing is on average 12 basis points as per the median and 27 basis points as per the mean, which is about 35 times more than the fees. This confirms the argument presented by market practitioners whom the author interviewed, suggesting that the biggest friction for the primary market transactions is the large size of the creation unit block and the often hard-to-collect

Figure 4.2: Median absolute mispricing by number of APs

The graphs below shows the median level of absolute relative mispricing computed across ETFs with different number of active APs and registered APs. Data span daily observations covered in N-CEN filings filed between July 2018 - August 2021. Reports that cover the period shorter than a year were excluded.



(a) Average mispricing (% over nav) by number of active APs

(b) Average mispricing (% over nav) by number of registered APs

Source: Author's computations.

components of the creation baskets rather than the fees charged by the ETF issuer, which are rather negligible for the arbitrage incentives. Another interesting statistic presented in this table is the distribution for estimated potential maximal arbitrage profit if one creation unit were to be transacted at the end-of-day prices when taking into consideration the creation fees charged by the ETF issuer and disregarding any other transaction cost (e.g., price impact). Here, we can see that potential arbitrage profits are rather small most of the time, with a median value of $2,200. This may partially explain why the zero-fund-flow days are as frequent as four days out of the business week on average.

The list of APs that appeared in the database is included in the appendix, Table A.2. The parsed database includes a total of 58 unique APs, with only 42 showing any primary market activity in any ETF. Figure A.1 shows the ranking of the APs as measured by the total primary market volume transacted in the collected database. Figure A.2 shows the ranking of the APs as measured by how frequently they appear as the first (biggest) AP per ETF report. We can notice that the biggest players in this market are mostly the biggest banking groups in the US with the largest assets under management, for example, Bank

Table 4.2: Creation Fees, summary statistics

This table presents the summary statistics for creation fees charged by the ETFs. The observation unit are daily data spanning the days covered in N-CEN filings submitted July 2018 — August 2021. Creation Fee reported below follows the convention from ETF Global database (with fees charged as the percent of dollar value creation not included). Max Profits defined as the difference between absolute (end of day) mispricing and creation fee.

	count	mean	sd	p1	p5	p10	p25	p50	p75	p90	p95	p99
Creation Fee [$ '000]	1,520,693	1.4	2.6	0	.15	.25	.25	.5	1	3.2	6	15
Creation Fee [bp of 1CU]	1,361,710	.7	1.1	0	.046	.076	.16	.35	.69	1.6	2.7	6
Absolute mispricing [bp of NAV]	1,458,260	27	54	0	.78	1.7	4.5	12	30	64	98	220
Max Profit [bp]	1,361,385	26	46	0	.4	1.3	4	11	29	62	96	215
Max Profit for 1CU [$ '000]	1,363,635	6.3	17	0	.065	.24	.79	2.2	5.9	14	24	64
Creation Unit Size [$ mln]	1,363,960	2.7	2.9	.38	.65	.86	1.2	1.8	3	5.4	8	14
Creation Unit Size as % of AUM	1,348,626	7.3	16	.014	.046	.096	.35	1.5	5.9	20	33	100

Source: Author's computations

of America, Goldman Sachs, JP Morgan, and Morgan Stanley. However, a few smaller institutions specializing in ETF market making are high on this list too, including Virtu Financials and Jane Street, as are some European financial institutions like ABN Amro, Société Générale, Credit Suisse, and Deutsche Bank.

Additional summary statistics on the creation fees charged by the ETF issuers are included in the internet appendix. Table A.4 shows the qualitative distribution of the types of fees charged by the ETFs for creations and redemptions. By far, the most common fee type is a fee charged per order. This is the type of fee that you can find being reported in Bloomberg or ETF Global. Table 4.2, however, shows that sometimes, ETFs also charge fees calculated per creation unit or per dollar created. Also, N-CEN filings suggest that in some cases, ETFs charge a combination of fees. A large number of cases with zero fees charged (triple *No* cell) could be the result of either no creations or redemptions during the reported period, zero fees charged, or data not provided; thus, they should be interpreted with caution. Table A.5 shows the quantitative distribution for the above-mentioned types of fees charged. As we can see, for the majority of ETFs, the creation fees seem very marginal, especially given the ETF creation unit sizes. Generally, they tend to be slightly higher for redemptions and in-cash transactions.

The summary statistics for the ETF primary market structure presented so far were for the whole universe of the ETFs registered in the US under the Investment Company Act of 1940 without looking at any cross-sectional difference that could arise across different categories. Thus, additional figures for category splits are included in the internet appendix. Figure A.3 shows the most relevant primary market statistics by fund size quantile, Figure A.4 shows relevant statistics by fund class, and Figure A.5 shows relevant statistics by fund age. Here, one can easily see that the ETFs that are most susceptible to poor competition structure in their primary market are smaller and younger. The average HHI for fixed-income and multi-asset class ETFs is generally higher than for the equity class. However, there is quite a large dispersion in market concentration in both of the ETF classes. The real estate ETFs' market concentration index appears to be comparable to the equity ETFs' levels. Due to the small representation of commodities and currency ETFs in the created database, I refrain from inferring about the whole commodity and currency ETF space based on the results obtained.[17]

Partial correlations for creation unit size and creation fees with fund characteristics are presented in Table 4.3. The creation unit tends to be higher the older and the bigger the fund is. That is most likely related to the fact that newly created and small funds tend to choose a smaller creation unit size to encourage liquidity and inflows. Also, funds with a larger pool of securities in their portfolio tend to have larger creation unit sizes. As liquidity and growing the assets under management are no longer priorities, funds may prefer larger creation unit sizes to lower the administration costs, lower the cash component of the creation baskets, and improve the benchmark tracking, if a fund is index-tracking. Another interesting result is that the creation unit size for fixed-income ETFs is on average higher than for the equity ETF by $0.6 million. This is most likely due to the less favorable treatment of odd lot trades in the fixed-income sphere and the relatively high minimal principal amount for a single bond. Potentially, one of the consequences of the advent of the fixed-income ETF trend could be a push towards lowering the average transaction sizes in the fixed-income markets as the ETF creation baskets tend to be highly diversified.

Effect of ETF primary market concentration on ETF price efficiency

To evaluate the effect of market concentration in ETF primary market on ETF pricing efficiency, in Table 4.4, I regress the daily deviations of ETF prices from NAV on the market

[17] Most of the commodities and currency ETFs are actually commodity and currency pools that trade exclusively with the use of commodity and currency futures and derivatives. Because these holdings are not considered securities, those funds do not need to register as investment companies under the Investment Company Act of 1940, which is regulated by the Securities and Exchange Commission, and are thus not required to file N-CEN. Pure commodities and currency exchange traded products are required to register with the Commodities Futures Trading Commission and the National Futures Association under the Commodity Exchange Act ("CEA") instead. Naturally, any exchange-traded product, whether it is an Exchange Traded Fund, Exchange Traded Commodity Pool, Exchange Traded Instrument (synthetic or levered fund) or Exchange Traded Note, needs to register with the Securities and Exchange Commission under the Securities Act of 1933. However, only funds that invest in a combination of futures and securities need to register under the Investment Company Act of 1940.

Table 4.3: Creation Unit size and fund characteristics partial correlations

This table reports the results of regressing fund characteristics on creation unit size. Data are annual with average values during a year used.

	Creation Unit ($ mln)
	(1)
Age (years)	0.2***
	(0.006)
AUM ($ bln)	0.07***
	(0.003)
# of holdings ('000)	0.8***
	(0.04)
Asset class:	
Equity (reference category)	0
	(.)
Fixed Income	0.5***
	(0.08
Multi Asset	-0.5**
	(0.2)
Real Estate	-0.4*
	(0.2)
Constant	1.1***
	(0.05)
Observations	6044
r2	0.3

Standard errors in parentheses
* p<0.10 ** p<0.05 *** p<0.01

Source: Author's computations.

concentration measures, namely, number of registered APs, the number of active APs, and HHI. To eliminate the variation in the market structure due to the specificity of the ETF, I include the ETF fixed effect along with a few time-varying control variables. Such an approach allows me to exploit within-ETF time variation in the primary market concentration measures to identify the effect of the competition structure on the level of mispricing. Significant negative coefficient estimates for the number of registered APs and the number of active APs confirm that having more APs present in the primary market can help to lower the magnitude of mispricings. The coefficient estimate in front of the primary market HHI is positive and statistically significant in almost all specifications as well, supporting the argument that the less competitive the primary market is, the higher the mispricings are on average.

Table 4.4: Effect of primary market structure on ETF price deviations from NAV, within ETF variation

This table presents the results of fixed effect regressions of relative mispricing relative to NAV on primary market structure measures such as number of registered APs, active APs and HHI index computed using primary market transactions volumes on the absolute mispricing. Sample includes ETFs older than 1 year and with HHI index above 0.33. Levered ETFs are excluded. Dates for which the reporting period of the N-CEN was less than a year are excluded. Regressions with constituents bid ask spread, columns (4)-(6), contains ETFs for which at least 95% constituents (as measured by weight) were matched with CRSP stock data. Observations are daily. Number of registered APs, active APs and HHI index is from the ETF's N-CEN filing with reporting period that overlapped with the daily observation. Data span the reporting periods for the N-CEN filings filed by the ETF between July 2018 — August 2021, thus the starting dates included in the sample are July 2017 for ETFs with fiscal year ending in June and later for fiscal year ending in other months.

	Relative Asbolute Mispricing (over NAV, in basis points)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of registered APs	-0.6***			-0.6***	-0.9***			-0.8***
	(0.06)			(0.06)	(0.1)			(0.1)
Number of active APs		-1.1***		-0.7**		-1.8***		-0.9*
		(0.3)		(0.3)		(0.6)		(0.5)
HHI of primary market (volume)			4.4**	1.1			9.0**	3.9
			(1.9)	(1.9)			(3.5)	(3.4)
Creation Unit [mln $ previous month]	1.4***	1.6***	1.5***	1.4***	1.5*	2.0***	2.2***	1.5*
	(0.4)	(0.4)	(0.4)	(0.4)	(0.8)	(0.7)	(0.7)	(0.9)
Creation fee [bp of 1CU previous month]	0.8	0.1	0.2	0.7	4.9	3.1	3.3	4.5
	(1.3)	(1.3)	(1.3)	(1.3)	(3.1)	(3.0)	(3.0)	(3.0)
AUM [bln $ previous month]	-0.4*	-0.6***	-0.6***	-0.4**	-0.9**	-1.2***	-1.2***	-0.9**
	(0.2)	(0.2)	(0.2)	(0.2)	(0.4)	(0.4)	(0.4)	(0.4)
Volume [mln $ previous month]	0.003	0.003	0.003	0.003	0.01	0.01	0.01	0.01
	(0.002)	(0.002)	(0.002)	(0.002)	(0.009)	(0.009)	(0.009)	(0.009)
Bid Ask Spread [% previous day]	12.8***	12.8***	12.9***	12.8***	10.9***	11.0***	11.1***	10.9***
	(1.3)	(1.3)	(1.3)	(1.3)	(1.7)	(1.7)	(1.7)	(1.7)
Vix	1.1***	1.1***	1.1***	1.1***	0.9***	0.9***	0.9***	0.9***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Constituents Bid Ask Spread [% previous day]					3.7*	3.8*	3.8*	3.8*
					(2.1)	(2.2)	(2.2)	(2.1)
N	754288	754288	754288	754288	237959	237959	237959	237959
Num of ETFs	1693	1693	1693	1693	651	651	651	651
R squared	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
ETF fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SE clustered at the ETF level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses; * $p<0.1$ ** $p<0.05$ *** $p<0.01$

Source: Author's computations.

Since the bigger, popular funds tend to have a more competitive primary market structure, while smaller funds are more prone to suffer from a less populated primary market, in Table 4.5, I report the coefficient estimates estimated separately for five subgroups of ETFs that were split by quintiles in the assets under management. We can see a clear pattern of high magnitudes in coefficients for the lowest size quintile funds, and this effect diminishes for larger ETFs. In the most affected subgroup, the smallest quintile, activating an additional AP decreases the level of mispricing by 2.6 basis points, and in the equity subsample, this number is almost twice as large.

It is important to note that the specification with the ETF fixed effect still suffers from the endogeneity problem, in particular the simultaneity bias. It is highly possible that an inactive AP will decide to become active if it observes high magnitudes of mispricing (or a non-registered financial institution will decide to register as an AP). This simultaneity causes the coefficient in front of the number of active (or registered) APs to be biased. However, because the two effects are acting against each other, the parameter estimates are biased towards zero (or rather towards the positive region); thus, we can interpret them as the lower bound on the causal effect of the number of active APs on ETF price efficiency.[18] This means that the true causal effect of the number of active APs on ETF price efficiency has even higher magnitudes.

Table A.6 in the appendix, shows the basic summary statistics of daily variables used in regression in Table 4.4, in which I restricted the sample to non-levered funds that are at least 1 year old and have an HHI of 0.33 or higher. We can see the ETFs traded at premiums more often than at discounts, though the distribution of mispricing is close to being symmetric. Also, the average magnitude of the discount is about \$0.05 in absolute terms, or 15 basis points in relative terms. Unlike in the secondary market, the primary market transactions do not happen often, on average every fifth day we observe a fund inflow or outflow to the

[18]If the endogeneity is of the simultaneity type, we can easily interpret the sign of the bias even in multivariate regressions. A working paper by
citetBasu provides the explanation for the sign of the asymptotic bias in the case of bi-directional causal effects.

Table 4.5: Effect of the primary market structure by size quintile

This table presents the results of fixed effect regressions of absolute mispricing relative to NAV on market structure measures such as number of registered APs, active APs and HHI index. The effect of the market structure on the level of mispricing was estimated separately for size quintiles by interacting the relevant variable with quintile dummies. Quintiles were assigned a month before a daily observation of mispricing. Sample includes ETFs older than 1 year. Levered ETFs are excluded. Dates for which the reporting period of the N-CEN was less than a year are excluded. Regressions with constituents bid ask spread, columns (4)-(6), contains ETFs for which at least 95% constituents (as measured by weight) were matched with CRSP stock data. Observations are daily. Number of registered APs, active APs and HHI index is from the ETF's N-CEN filing with reporting period that overlapped with the daily observation. Data span the reporting periods for the N-CEN filings filed by the ETF between July 2018 — August 2021, thus the starting dates included in the sample are July 2017 for ETFs with fiscal year ending in June and later for fiscal year ending in other months.

Dependent Variable:	Relative Asbolute Mispricing (over NAV, in basis points)					
	(1)	(2)	(3)	(4)	(5)	(6)
Size dummy interacted with:	*# registered APs*	*# active APs*	HHI index	*# registered APs*	*# active APs*	HHI index
Size Q1	-0.6***	-2.6***	8.2*	-0.7***	-4.2**	15.2**
	(0.1)	(1.0)	(4.2)	(0.2)	(1.9)	(7.4)
Size Q2	-0.7***	-1.1***	2.9	-0.7***	-1.0*	2.6
	(0.06)	(0.4)	(2.3)	(0.1)	(0.6)	(3.9)
Size Q3	-0.5***	-0.4	-1.2	-0.6***	-0.8**	1.5
	(0.05)	(0.2)	(2.1)	(0.08)	(0.3)	(3.3)
Size Q4	-0.4***	-0.4**	7.8***	-0.4***	-0.9***	10.5***
	(0.04)	(0.2)	(1.8)	(0.07)	(0.3)	(3.3)
Size Q5	-0.3***	0.04	8.1***	-0.3***	-0.2	7.2**
	(0.03)	(0.1)	(1.9)	(0.05)	(0.2)	(2.9)
N	1091692	1091692	1091692	360912	360912	360912
Num of ETFs	1933	1933	1933	752	752	752
R squared	0.3	0.3	0.3	0.3	0.3	0.3
SE clustered at the ETF level	Yes	Yes	Yes	Yes	Yes	Yes
Control Variables:	ETF fixed effect, Creation Unit Size, Creation Fee, Size Quintile FE, ETF daily volume ($), ETF Bid Ask Spread, Vix			ETF fixed effect, Creation Unit Size, Creation Fee, Size Quintile FE, ETF daily volume ($), ETF Bid Ask Spread, Vix, Weighted Constituents Bid Ask Spread		

Standard errors in parentheses
* p<0.10 ** p<0.05 *** p<0.01

Source: Author's computations.

ETF. Creation unit size is on average \$2.6 million while the fund size is on average \$1.1 billion.

To address the potential concerns about a spurious relationship between the market concentration measure and the level of observed mispricing due to time trends, Table A.7 in the appendix reports estimation results with a time fixed effect included in the form of monthly dummies. The majority of estimates in front of the market concentration measures remained significant, with only slightly lower magnitudes.

So far, all the effects were estimated as an average effect among all ETFs with a not-so-competitive primary market, as defined by an HHI larger than 0.33. However, one could wonder how the effect differs across different starting points in the primary market concentration. Thus, Figure 4.3 shows the semi-elasticities of having an additional AP transacting for different initial levels of primary market structure. Naturally, the less competitive the initial primary market is, the stronger the effect is of having an additional AP transacting in it. In particular, going from zero to one transacting AP corresponds to a 7% decrease in the level of mispricing in the equity ETFs. Also, interestingly, in the case of the equity ETFs, even at a relatively high number of active APs, up to 9, the effect still appears significantly different from zero, although much smaller in magnitude.

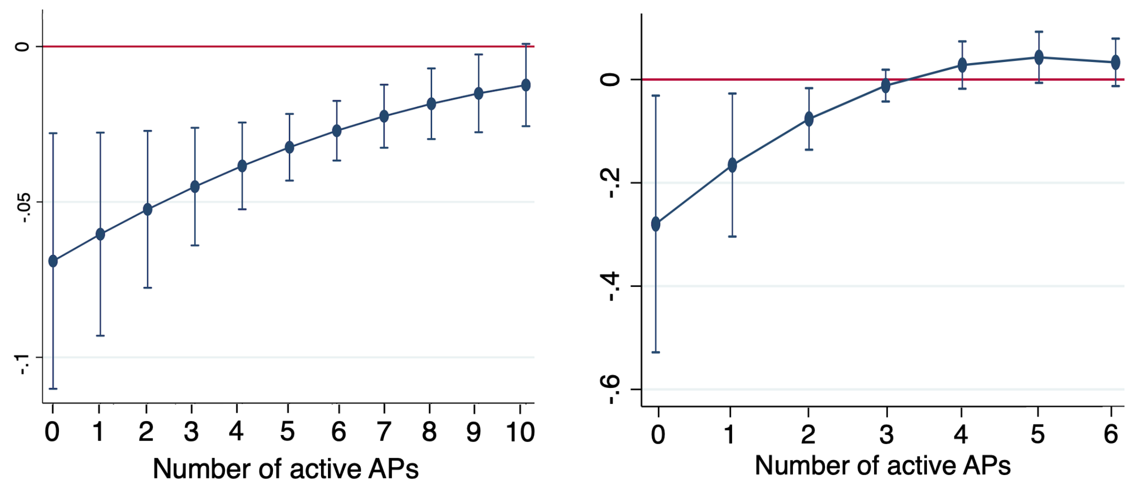
Figure 4.3: Semi-elasticities of absolute relative mispricing to number of active APs.

Semi-elasticities obtained by estimating a regression separately for equity ETFs and for fixed income ETFs with a third order degree polynomial effect of the number of active authorized participants, ETF fixed effect and time varying controls as in column (2) of Table 4.4.

$$Ln(AbsRelativeMispricing_{it}) = \beta_1 NumActiveAP_{it} + \beta_2 NumActiveAP_{it}^2 + \beta_3 NumActiveAP_{it}^3$$
$$+ \sum_{k=1}^{K} \gamma_k TimeVaryingControl_{k,it} + \alpha_i + \epsilon_{it}$$

Standard errors are clustered at the ETF level. Semi-elasticities computed at the mean values. Bars represent 90% Confidence Intervals.

(a) Equity ETFs (b) Fixed Income ETFs



Source: Author's computations.

ETF Pair Analysis

Another empirical strategy to test the hypothesis about the relevance of the primary market structure to the level of mispricing is to look at the differences in potentially similar ETFs. Here, I turn back to the PQ model of the ETFs introduced earlier in this paper, as it has some important empirical implications. In particular, the more elastic the supply curve is (for example, as a result of a higher number of APs), the smaller the price fluctuations should be, and the larger the shares' fluctuations should be. To test this implication, I employ the pair analysis, where I construct ETF pairs that follow the same benchmark index. Presumably, ETFs following the same benchmark should be subject to similar market conditions, creation baskets, and demand shocks.

Figure 4.4 shows graphically the relationships among the relevant within—ETF pair differences. We can see that the simple directional relationships, as portrayed by the fitted lines, support the PQ model's theoretical implications. ETFs with a higher number of active APs within the same index benchmarking pair have less volatile mispricings and more volatile fund flows. Similarly, ETFs with larger HHI levels relative to their benchmark-tracking paired ETF have more volatile mispricings and less volatile fund flows.

Although it would be easy to think about two ETFs following the same index as very similar to one another, there might be some important differences between them that could affect the difference in the observed level of mispricing and fund flows. Thus, in Table 4.6, I expand the simple regression specifications of the fitted lines to include the differences in fund size and creation unit size as well. We can see that the results still hold. The ETF with the larger number of active APs has less volatile mispricings and more volatile fund flows, and the ETF with the higher market concentration has more volatile mispricing (as per sign, although the coefficient is not statistically significant) and less volatile fund flows. Also, ETFs with a higher number of APs had on average more frequent fund flows.

To address the concerns that the relationship between ETF issuer and AP can be a driving force in the results, table A.9 in the appendix presents similar results when controlling for the ETF pair fixed effect. We can see that even in this case the results mostly hold with just a slightly smaller point estimate.

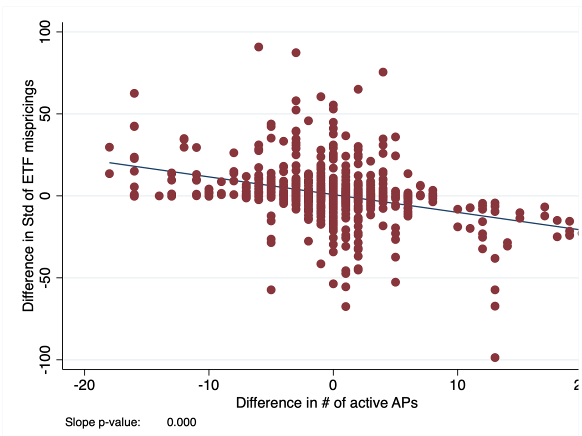
Effect of the Creation Unit Size on ETF price efficiency

As mentioned earlier, one of the important frictions pointed out by the market practitioners that makes arbitraging in the ETF primary market a non-trivial job is the feature of the creation unit size blocks, especially when those blocks are large. There is very little variation in the creation unit size across ETFs as measured by the number of shares, and ETFs typically set the creation unit size at 50,000 shares. US ETFs' regulatory policies prior to

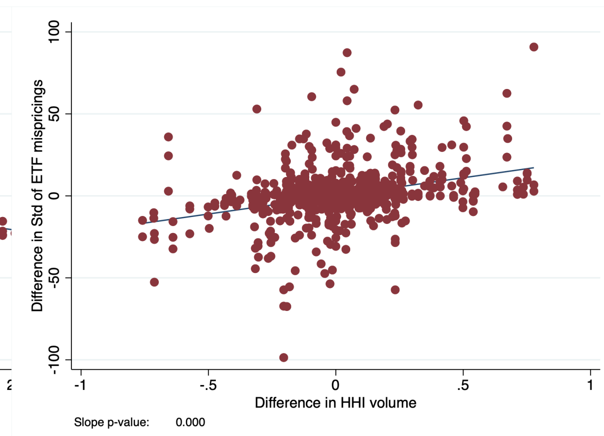
Figure 4.4: ETF pairs - scatter plots

Figure below presents the scatter plots for the difference in number of active APs or difference in primary market HHI index plotted against the difference in the volatility of mispricings or volatility of fund flows. The differences are computed between the two ETFs constituting one ETF pair, where the ETF pair is constructed such that the two ETFs have the same primary benchmark index in ETF Global. Only non-levered ETFs are considered. Each dot represent a quarteyl data. Standard deviations are computed over each quarter using daily observations. Data spans 2017 Q3 - 2021 Q2. Mispricing measure used in this exercise is relative mispricing over NAV, in basis points. Fund flow variable used in this exercise is daily fund flows in mln dollars with zero fund flow days included. In total, there are 89 ETF pairs included over 53 different primary benchmarks.

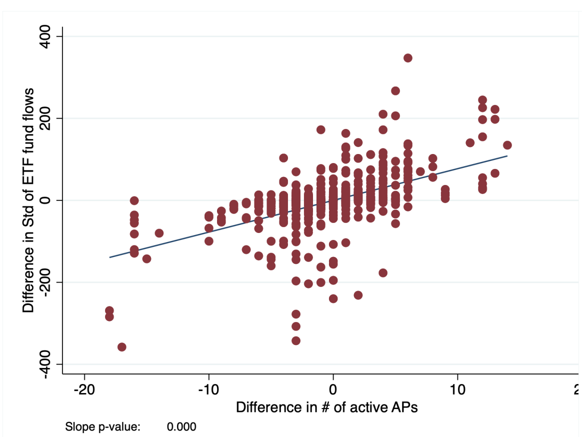
(a) Mispricings on # of active APs



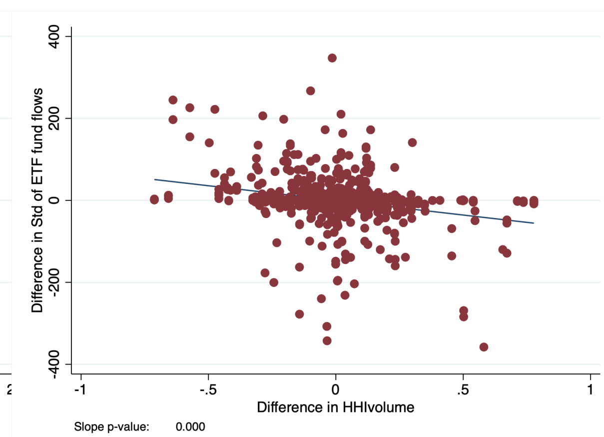
(b) Mispricings on HHI



(c) Fund flows on # of active APs



(d) Fund flows on HHI



Source: Author's computations.

Table 4.6: ETF Pairs Analysis - regressions

This table reports the results of the regressions for ETF pairs. ETF pairs were constructed as two ETFs with the same benchmark index. Unit of observation is the difference or log-difference of characteristics between two ETFs within one ETF pair. Only non-levered ETFs are considered. Standard deviations and the frequency of non-zero fund flow days are computed over each quarter using daily observations. Regressors are computed as median observation each quarter. Data spans 2017 Q3 - 2021 Q2.

	Diff in StD of Relative Mispricing (in basis points)		Diff in StD of fund flow (in basis points)		Diff in frequency of days with nonzero fundflows (in %)	
	(1)	(2)	(3)	(4)	(5)	(6)
Diff in # active APs	-0.6**		6.1**		1.9***	
	(0.2)		(2.6)		(0.4)	
Diff in HHI_volume		19.1***		-162.6***		-22.0***
		(6.3)		(43.5)		(6.7)
ln(Ratio of Creation Unit Size in Dollars)	2.8	2.1	-19.1	-12.4	-12.3***	-11.1***
	(2.2)	(2.1)	(14.5)	(14.0)	(2.3)	(2.3)
ln(Ratio of AUM)	-1.9**	-1.7**	-21.8***	-22.1***	9.4***	10.7***
	(0.9)	(0.7)	(6.7)	(5.7)	(1.0)	(0.8)
Constant	0.3	0.4	3.7	3.0	-2.3*	-2.6*
	(1.3)	(1.3)	(9.4)	(9.5)	(1.3)	(1.4)
Observations	759	759	697	697	740	740
r2	0.1	0.1	0.1	0.1	0.7	0.7
SE clustered at:	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair
# of ETF pairs	85	85	82	82	83	83

Standard errors in parentheses

* $p<0.1$ ** $p<0.05$ *** $p<0.01$

Source: Author's computations.

2018 are largely at the heart this, as there was a minimum size requirement for the creation unit size of 50,000 shares or \$1 million. Only in 2017 and 2018, was this requirement eliminated, giving ETFs more freedom in choosing the preferred size of their creation unit.

In the data, creation unit size adjustments are rarely observed throughout the ETF lifespan, and if such adjustments happen, they usually take place around ETF splits or reverse splits. To the best of my knowledge, so far there have not been very many studies or even discussions about the effects that the creation unit size may have on ETF pricing efficiency. However, in one of the SEC notices,[19] we can read that *"a reduction in the size of a creation unit may provide potential benefits to investors by facilitating additional creation and redemption activity ... thereby potentially resulting in increased secondary market trading activity, tighter bid—ask spreads and narrower premiums or discounts to NAV."* Thus, it seems very natural to investigate to what extent the creation unit size contributes to ETF pricing efficiency.

Table 4.7 shows the distribution of the number of created or redeemed units. Daily fund flows data are used to construct this table; thus, it shows the number of created or redeemed units within one day and not within one transaction. However, multiple APs could submit creation or redemption orders on the same date. If such a table were created for each creation/redemption order, the distribution would be even more concentrated towards a smaller number of units. Thus, we can immediately see that at least 30% of all creation redemption orders are transacting at only one creation unit.

ETFs, similarly to stocks, may undergo splits or reverse splits. For splits, the typical reason behind such action is to lower the price, making it more accessible to retail investors. Reverse splits are less frequent and usually happen for ETFs where the price has fallen too much, which is often the case for synthetic and levered ETFs. Arguments made to justify a reverse split could be to avoid higher transaction fees, risking the "junk" ETF label, or being delisted from the stock exchange. Sometimes, ETF splits and reverse splits are followed by a change

[19] SEC's letter as of July 16, 2015, available at: https://www.sec.gov/rules/sro/nysearca/2015/34-75475.pdf

Table 4.7: Created Units

Unit of observation is daily fund flow to the ETF. Data span the reporting periods for the N-CEN filings filed by the ETFs between July 2018 - August 2021, thus the starting dates included in the sample are July 2017 for ETFs with fiscal year ending in June and later for fiscal year ending in other months. Creation Unit Size used to generate the table is the minimum between creation unit size in ETF Global database and creation unit size reported in the N-CEN filing for relevant period. In case the created units are not an integer, the number is rounded to the nearest integer. For Vanguard's ETFs however, due to their policy of allowing mutual funds shares conversion, the created units are rounded to the floor integer.

	% of all redemptions (creations) days	
	Full Sample	ETFs with HHI volume>0.33
Redeemed Units		
>50	4%	2%
11-50	17%	12%
4-10	27%	26%
2-3	26%	28%
1	26%	31%
Created Units		
1	28%	32%
2-3	29%	31%
4-10	27%	25%
11-50	14%	10%
>50	3%	1%
No fund flows	% of all days	% of all days
0	83%	89%

Source: Author's computations.

in the creation unit size to keep it at roughly the same level; however, most of the time, they are not. In such a case, the ETF split effectively significantly changes the dollar value of the creation unit, although keeping the number of shares constant.

Table 4.8 shows the event study around the decreases in the creation unit size of the ETF or around ETF splits that were not accompanied by other events of this nature. Reverse splits and creation unit size increases are not included due to their mere presence and concentration in synthetic and levered ETFs. As we can see, there were more cases of unaccompanied ETF splits than unaccompanied creation unit size decreases. As we can see in columns (3) and (4), after the creation unit size is cut in half, mispricing decreases by 1.8 bp if the change

Table 4.8: Effect of ETF share split or decrease in creation unit size on mispricings

This table shows the effect of ETF splits (not accompanied by the simultaneous decrease in creation unit size) and of the decrease in Creation Unit Size (not accompanied by the ETF split) on the mispricing magnitudes and. Creation Unit Size change events are identified using daily data in creation unit size from ETF Global, which is available starting from year 2016. ETF split events are identified using the cumulative adjustment factor for shares (cfacshr) in CRSP. Magnitude variable represents the magnitude of the split or decrease and is larger than 1. Daily observations around the event are included from 1 to 90 days prior and post event. Leveraged and volatility focused ETFs are excluded. Only ETFs older than 1 year are included. Dummy variable equals 0 prior to the event and 1 after the event. Dummy variable equals 0 prior to the event and 1 after the event. Mispricing computed at the end of day.

	Absolute Mispricing (in $)		Absolute Relative Mispricing (over NAV, in bp)	
	(1)	(2)	(3)	(4)
dummyDecrease # ln(Magnitude)	-0.006		-2.6**	
	(0.004)		(1.2)	
dummySplit # ln(Magnitude)		-0.2***		-2.1**
		(0.03)		(0.9)
Observations	3852	13505	3852	13505
R-squared	0.419	0.461	0.370	0.568
Fixed Effect	Event	Event	Event	Event
SE clustered at event level	Event x YearMonth	Event	Event x YearMonth	Event
# of events	39	158	39	158
Events starting date	2016	2011	2016	2011

Standard errors in parentheses

* p<0.1 ** p<0.05 *** p<0.01

Source: Author's computations.

was in the creation unit size block and by 1.5 bp if the change was due to a split.[20]

The fact that the magnitude of the split effect after the split is slightly lower is most likely due to other channels of split events that affect share prices. One of such channels often mentioned is the increase in demand due to the split. Table A.10 in the appendix addresses this issue by estimating the effect of such events on the level of premiums and discounts separately. Here, we can see that the split effects are mostly concentrated in decreasing the levels of discounts rather than premiums. This is consistent with the increased demand story, when the prices of split ETFs are pushed upwards. Interestingly, the opposite seems to happen after a decrease in the creation unit size block. This could be because the premiums seem to be higher than the discounts prior to the decrease event. However, given the small number of observations, we should interpret those differences with caution.

[20] Halving the creation unit size corresponds to a magnitude of 2; thus, the coefficient reported needs to be multiplied by ln(2) for such an effect estimate.

CHAPTER 5

Model

To better characterize the effect that the competition and the creation unit have on the ETF primary market, I constructed an intertemporal choice model for the AP. There are two crucial mechanisms that capture the AP's behavior. First, there is a value option of waiting for the mispricing to widen. This means that instead of fully utilizing today's mispricing, the AP may have an incentive not to exercise a mispricing today at all, or to exercise at a smaller scale in order to secure better arbitrage opportunities tomorrow. This is particularly salient in the monopolistic competition case. The second important ingredient of the model is the introduction of the competition. The effect of improving the competition in the primary market would be such that not only would an additional AP transact on it, but also its presence would discourage the first AP from waiting for the mispricing to widen because there is a risk that another AP would exercise a mispricing today and thus lessen the arbitrage opportunities tomorrow.

The empirical analysis of the ETF primary market structure hinted at another important feature of this market, in particular, the asymmetry between APs. The HHI is 0.5 on average. In the symmetric market structure, this would correspond to a duopoly case. However, as we have seen, the average number of active APs is twice as high. This suggests that APs in the individual ETF primary market are not symmetric in the sense that they differ significantly in transaction volumes. This observed feature prompted me to propose a model with heterogeneous AP, where agents are heterogeneous in the transaction costs and outside opportunities for their capital.

As the creation unit size seems to be a relevant aspect of this market, and often a relevant friction, I also equip the model with this discreteness feature of creation and redemptions. This not only better captures the primary market structure with lumpy transactions but also allows for analysis of the creation unit size policy.

In the subsections below, I build up the model from the simplest static monopoly case by adding each individual component to it step by step, arriving at the final model specification used, which is later calibrated to match relevant data moments.

5.1. General specification

The authorized participant (agent i) chooses $x_{it} \in R$, which denotes how much to create/redeem at time t after observing u_t. The state variable u_t denotes the deviation of an ETF's price from the value of its constituents (or NAV). Naturally, if the deviation of the ETF price is positive (negative), it should (probably) only consider creation (redemption) which corresponds to positive (negative) x_{it}. Upon creation/redemption, the authorized participant gets profit $u_t x_{it}$ minus the transaction costs $\lambda^i x_{it}^2$.

The mispricing process follows a mean-reversion process with an endogenous autoregression parameter that depends on two elements: first, parameter $\rho \in (0,1)$, which captures the mean reversion of the mispricing process even in case of permanent inactivity of authorized participants, which corresponds to the fact that other participants in the market may engage in "arbitrage-like" transactions; second, the reversion of u_t depends on the price impact of AP creation/redemption activity on the deviation of the ETF price denoted by function $\psi(.)$. The function $\psi(.)$ should take values in $(0,1]$ and sign($\psi'(X)$) = - sign(X), and $\psi(0) = 1$.[21] The liquidity shocks (or demand shocks) to the ETF are captured by ϵ_{it}. It is worth noting that in this specification, the low of motion for the state variable in this specification is endogenous and, as such, will be an equilibrium element. Also, to better capture the structure of the ETF primary market, I introduce the creation unit size constraint.[22]

$$V_i = \max_{x_{it}} \sum_{t=1}^{T} \beta^{t-1} E[v(CF_{it})]$$

[21] Function $\psi(.)$ could have, for example, a bell shape. The Gaussian function of form $\psi(x) = \exp^{-x^2/\phi}$ with parameter $\phi > 0$ representing the speed of price impact could be employed.

[22] In reality, the AP can only create and redeem whole numbers of creation units. Thus, the proper specification includes the grid for creation/redemption with an interval equal to the creation unit size. However, to simplify analytical considerations, I will first consider a continuous domain for the level of creation/redemption. Later, in the full specification version of the model, I introduce the grid optimization with one grid unit corresponding to one creation unit

where

$$CF_{it}(x_{it}) = \begin{cases} 0 & \text{if } x_{it} = 0 \\ u_t x_{it} - \lambda^i x_{it}^2 - f^i & \text{if } x_{it} \neq 0 \end{cases}$$

s.t.

$$u_{t+1} = \rho\psi\left(\sum_i x_{it}\right)u_t + \epsilon_{t+1} \qquad \epsilon_t \sim N(0, \sigma^2)$$

$$u_0 = 0$$

$$x_{it}u_t \geq 0 \qquad (nSC)$$

$$|x_{it}| \in \{0cu, 1cu, 2cu, 3cu, ...\} \quad \textit{(cu - creation unit size)}$$

The nSC constraint (non-speculator constraint) captures the fact that APs cannot act in a way that could contribute to even bigger ETF mispricing tomorrow. In particular, APs are not allowed to create when an ETF is traded at a discount nor to redeem if an ETF is traded at a premium. However, this constraint is redundant with the choice price impact function $\psi(.)$ that I added later. The concept of a fixed transaction cost per creation/redemption order is captured by a parameter f^i in the CF function. However, in my later considerations, I will assume that the fixed cost equals zero.

5.2. Model considerations

To simplify the theoretical considerations below I assume that the fixed cost is zero, $f = 0$, AP is risk neutral, $v()$ is linear, and the AP can create or redeem ETFs at any level of creation $x \in R$.

Case 1a: N=1, T=1 (myopic monopoly)

This is the simplest myopic case to consider. There is no intertemporal choice in this case.

The optimal level of creation/redemption given the realization of u is

$$x^* = \frac{u}{2\lambda}$$

with the optimal level of CF to be

$$CF(u, x^*) = \frac{u^2}{4\lambda}$$

Case 2a: N=1, T=2 (two-period monopoly)

This is the extension of Case 1a, which captures the intertemporal choice consideration in the simplest possible way. The model can be rewritten as:

$$V(u) = max_{x_i, x_2} \quad CF_1(x_1) + \beta E\big(CF_2(x_2)\big)$$

where

$$CF_t(x_t) = u_t x_t - \lambda x_t^2 \qquad t \in \{1, 2\}$$

s.t.

$$u_2 = \rho\psi\big(x_1\big)u_1 + \epsilon \tag{5.1}$$

$$u_1 = u$$

$$x_t u_t \geq 0$$

This means that the action the AP takes in period 1 in the form of creation/redemption level x_1 will affect the distribution of ETF mispricing tomorrow, for example, $u_2 \sim N\left(\rho\psi\big(x_1\big)u_1, \sigma^2\right)$.

We could assume that the AP will decide every period how much to create/redeem. However, this equilibrium would be equivalent to deciding at period 1 about how much to create/redeem at both periods 1 and 2 as long as the optimal strategy for $t = 2$ will be a function of a shock at that period ϵ_2 (or mispricing u_2). Thus, we can easily plug the policy function for period 2 using the myopic solution and rewrite:

$$V_i = \max_{x_{i1}} \quad v(CF_{i1}(x_{i1})) + \beta E\Big[v(CF_{i2}(x^{*myopic}(u_2)))\Big]$$

s.t.

$$u_2 = \rho\psi\big(x_{i1}\big)u_1 + \epsilon_2$$

We can immediately see that the decision of how much to redeem today will affect not only the cash flow today, but also the distribution of mispricing tomorrow through the low of motion for u.

We then plug the IC from the myopic case straight into the value function as below.

$$V = \max_{x_1} \quad u_1 x_1 - \lambda x_1^2 + \beta E\Big(\frac{u_2^2}{4\lambda}\Big)$$

and optimize given the low of motion equation as in 5.1, recalling that $u_2 \sim N\Bigg(\rho\psi\big(x_1\big)u_1, \sigma^2\Bigg)$. Now, we just need to optimize over x_1.

$$E\Big(\frac{u_2^2}{4\lambda}\Big) = \frac{\big(\rho\psi\big(x_1\big)u_1\big)^2 + \sigma^2}{4\lambda}$$

Optimizing over x_1 gives FOC:

$$\big[u_1 - 2\lambda x_1\big] + \frac{2\beta\rho^2 u_1^2 \psi\big(x_1\big)\psi'\big(x_1\big)}{4\lambda} = 0$$

By rearranging we can write:

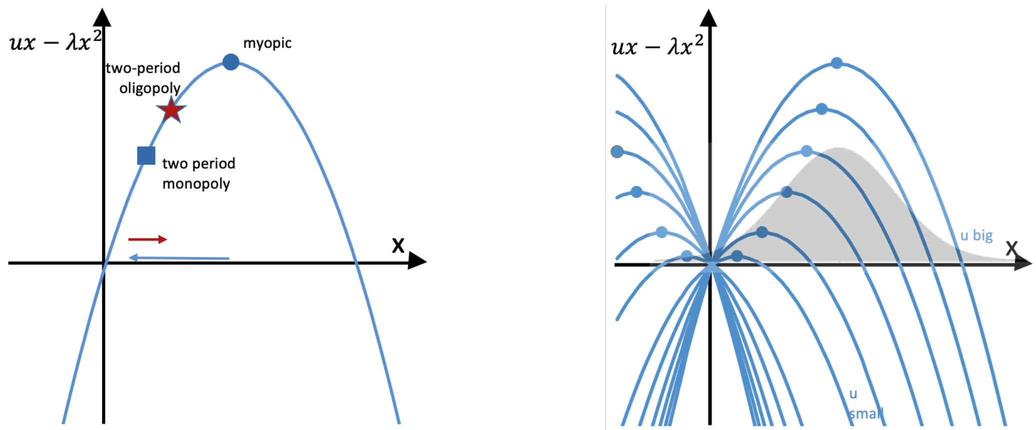
$$x_1 = \frac{u_1}{2\lambda}\Big[1 + \frac{\beta\rho^2 u_1 \psi(x_1)\psi'(x_1)}{2\lambda}\Big]$$

Recall that in the myopic case $x_1 = \frac{u_1}{2\lambda}$, so that if $u_1 > 0 (u_1 < 0)$, which means that if ETF is traded at a premium(discount), then $x_1 > 0$ ($x_1 < 0$), which means that AP would engage in creation (redemption). Because in equilibrium, $\psi'(x)$ would have the opposite sign to u, we can see that the magnitudes at which the AP would engage in creation and redemption in this two-period model at period 1 would be smaller than if it were myopic, which illustrate this intertemporal tradeoff between enjoying full arbitrage profits today versus securing yourself better arbitrage opportunities tomorrow. The graphical illustration of this effect on the choice of creation/redemption in the two-period model is shown in Figure 5.1a, where the optimal choice of x shifts to the left along the one-period profit curve, showing the option value of waiting in the intertemporal model. Figure 5.1b shows the possible realizations of tomorrow's mispricing (u) and the corresponding optimal choices of x and profits in the form of dots. Every time the AP creates or redeems when there is a premium or a discount, it shifts the distribution of tomorrow's mispricings (shaded area) towards zero. The less it creates/redeems, the less it shifts the distribution securing better arbitrage opportunities tomorrow.

Case 3a: N=1, T=∞ (Bellman equation, monopoly)

To cast the infinite-time problem into a Bellman equation, that is a contraction mapping, the cash flow function needs to be bounded. It would be natural to impose the minimum and maximum levels of possible creations/redemptions. In reality, you cannot redeem more shares than a particular ETF has issued. Also, the fact that there is a fixed and limited number of shares of the underlying stocks on the market provides a natural upper bound for possible ETF creations. Here, however, for analytical simplicity I assume that there are lower and upper bounds on the mispricing level $u \in [\underline{u}, \overline{u}]$. The ETF provider can

Figure 5.1: Theoretical model - illustration of AP optimization



(a) Myopic choice without minimum creation constraint

(b) Second period optimal strategy in two-period model with no minimum creation

actually temporarily suspend creation/redemption in certain limited circumstances (SEC permitting), and most likely, such suspensions would happen at high levels of mispricing. In this case, instead of assuming that ϵ has a normal distribution, we would rather assume that u has a truncated normal distribution.[23].

$$V(u) = \max_{x} CF(x,u) + \beta E[V(u')]$$

s.t.

$$\Big(u'|u,x\Big) \sim TN(\rho\psi\big(x\big)u, \sigma, \underline{u}, \overline{u})$$

The Lagrangian could be written as.

$$V(\mu,\epsilon) = \max_{x,\mu'} \quad F(\mu,x,\epsilon) + \beta E[V(\mu',\epsilon')] + \omega\Big(\rho\psi\big(x\big)u - \mu'\Big)$$

Assuming $F(\mu,x,\epsilon)$ is continuously differentiable in μ and the differentiability of a value function, the FOCs are:

[23]More technical details are provided in Appendix A.1.

$$x: \qquad \frac{dF}{dx} = -\omega\rho\psi'(x)u \tag{5.2}$$

$$\mu': \qquad \beta E\frac{dV(\mu',\epsilon')}{d\mu'} = \omega \tag{5.3}$$

Combining 5.2 and 5.3 we get the interior optimality condition:

$$\frac{dF(\mu,x^*,\epsilon)}{dx} = -\rho\psi'(x)u\beta E\left(\frac{dV(\mu'^*,\epsilon')}{d\mu'}\right) \tag{5.4}$$

where x^* is the optimal policy function for x which uniquely pins down the optimal next period μ'^*.

Using the stochastic equivalent of the Envelope Theorem and differentiating with respect to μ :

$$\frac{dV(\mu,\epsilon)}{d\mu} = \frac{dF(\mu,x^*,\epsilon)}{d\mu} \tag{5.5}$$

equation 5.4 becomes:

$$\frac{dF(\mu,x^*(\mu,\epsilon),\epsilon)}{dx} = -\rho\psi'(x)u\beta E\left(\frac{dF(\mu'(\mu,\epsilon),x^*(\mu',\epsilon'),\epsilon')}{d\mu'}\right) \tag{5.6}$$

An interesting observation is that equation 5.6 is effectively a stochastic Euler equation. It defines the optimality condition where the marginal cash flow loss from one less unit of creation today $-\frac{dF}{dx}$ must equal the discounted expected marginal cash flow benefit from the increased expected level of mispricing tomorrow $\beta E\frac{dF}{d\mu'}$, where the increased expected level of mispricing tomorrow is a marginal effect from the marginally less creation today $\rho\psi'(x)u$.

Additionally, assuming $F(\mu,x,\epsilon) = (\mu+\epsilon)x - \lambda x^2 = ux - \lambda x^2 = CF(u,x)$, we can write 5.6 as:

$$u - 2\lambda x^* = -\rho\psi'(x^*)u\beta E\left(\frac{dCF(u',x^*)}{du}\frac{du}{d\mu}\right) \tag{5.7}$$

which is similar to the two-period case if we rearrange it:

$$x^* = \frac{u}{2\lambda}\left[1 + 2\lambda\beta\rho\psi'(x^*)uE\left(\frac{dCF(u'(x^*), x^*(u'))}{d\mu}\right)\right]$$

Again, similarly to the two-period model, we can see that the magnitude of creation/redemptions will be smaller than in the myopic case. This is because $\psi'(x^*)u < 0$ and $\frac{dCF(u'(x^*),x^*(u'))}{d\mu} > 0$.

Case 1b: N>1, T=1 (myopic oligopoly)

If there are two or more myopic APs, they will not affect each other's choices at all. Their aggregate choice will, however, affect the low of motion for mispricing, but they won't realize it and thus won't internalize it in their decision-making. Also, because the aggregate level of creation/redemption of all APs does not affect the profits today but only affects tomorrow's mispricing level, the fact that there are other APs in the market will not change the optimal myopic choice of the AP. In general, the higher the number of N, the quicker the ETF mispricing would is reverted in simulations.

Case 2b: N>1, T=N (two-period oligopoly)

This is the case where things get interesting yet still solvable. Assuming that each individual AP knows about the existence of each other and knows each other's type characterized by (λ_i), they will now consider and internalize not only their own effect of creating/redeeming today on tomorrow's opportunity set, but also the optimal action taken by the other APs. We can incorporate this two-period oligopoly into a model by tweaking the $\psi(.)$ function into:

$$\psi(x_{i1} + X_{-i1})$$

where X_{-i1} denotes the optimal action taken by competing APs in the particular ETF market in period 1. This is an equivalent specification to the standard Cournot oligopoly, where firms compete with each other on quantities produced. Although this case has a flavor of dynamic games problems, its two-period specification helps to keep it somehow tractable yet allows for intertemporal effects of competition structure to be included.

In the particular case of the ETF primary market populated by APs that are symmetric in (λ^i), they would have the exact same value functions and policy functions, given mispricing u and competitors' action X_{-i1} today. It can be easily shown that the optimal creation/redemption policy in period 1 for each AP has to satisfy:[24]

$$x_{i1}^* = \frac{u}{2\lambda}\Big(1 + \frac{\beta\rho^2 u_1 \psi(Nx_{i1})\psi'(Nx_{i1})}{2\lambda}\Big) \tag{5.8}$$

We can again easily see that the optimal creation/redemption would be smaller in magnitude than the myopic case. This is because $u\psi'(.) < 0$ and $\psi(.) > 0$). If $N = 1$, equation 5.8 simplifies to the two-period monopoly case. As the competition among APs increases, thus as $N \to \infty$, x_{i1}^* approaches the myopic case.

Case 3b: N>1, T=∞ Bellman equation, oligopoly

The infinite horizon model with N>1 APs corresponds to the infinite-horizon stochastic dynamic games model. It can be formalized as a system of Bellman equations for all players and a transition function for the state variable:

$$V_i(u) = \max_{x_i} CF(x_i, u) + \beta E[V(u')|x_i, X_{(-i)}(u), u] \qquad \text{for} \quad i = 1, 2, ..., N \tag{5.9}$$

s.t.

$$\Big(u'|u, x\Big) \sim TN\bigg(\rho\psi\Big(\sum_i x_i\Big)u, \sigma, \underline{u}, \overline{u}\bigg)$$

[24]Technical details in Appendix A.3.

$$xu \geq 0$$

where X_{-i} is a vector containing the policy function of competing APs. $V_i(u)$ is the value function of agent i given that the competing agents remain committed to their policies $X_{(-i)}(u)$. The system of equations as in 5.9, with one such equation for each of the APs, $i = 1, ..., N$, defines a Markov perfect equilibrium (MPE). In my considerations, I will focus on pure-strategies equilibrium only. The model can be solved using the value function iteration algorithm similar to

citetPakesMcGuire, although some modifications are required. In particular, the state variable u in this model is assumed to be continuous. However, the value function iteration algorithm will operate on the discretized grid of u. Thus, at each iteration, we need to compute the discretization of the transition probabilities from state u today to any other state u' tomorrow on a grid $u^1, u^2, ...u^k, ...u^{Nu}$. We can assume that the discretization of transition probability would be computed in the following way:

$$Pr\big(u^j|u^k, x, X_{-i}\big) = P\Big(\frac{u^{j-1} + u^j}{2} < u' < \frac{u^j + u^{j+1}}{2}\Big|u^k, x, X_{-i}\Big)$$

A more detailed description of the algorithm used to solve the model is included in Appendix A.4.

Case 4: Model with freeze (symmetric)

Again, I begin by assuming that there are N APs in total. However, each AP may be hit with the freeze constraint with probability γ, which would stop this AP from engaging in creation/redemption today.

The introduction of the freeze constraint has multiple uses. On the one hand, it can strengthen the option value of waiting because the risk of competing APs creating/redeeming today is smaller. On the other hand, however, it causes another distortion, if an AP knows that it may be frozen tomorrow, it will not wait; thus, such a freezing friction can also lessen the effect of the intertemporal choice.

The freezing constraint can also help to create more realistic simulations and capture the fact that APs do not create or redeem shares every day. Alternatively, we can think about it as a chance of having an outside investment opportunity that would generate a higher return than the ETF creation/redemption or having other capital, technology, or attention constraints. Since an APs can be viewed as similar entities to market makers, in some regards, since many of the institutions engage in both roles across different ETFs, many of the considerations of traditional market making can be passed through to the ETF primary market. In particular, as noted in the paper by
citetYueshen, market makers' presence is uncertain over any short time interval and is subject to constraints, e.g., capital, technology, or attention. Similar considerations could be given to the arbitrageurs in the ETF primary market. However, the most striking difference between market makers and APs is that a market maker who takes on the role of a designated market maker is required to constantly quote and facilitate its buy and sell transactions, whereas APs' transactions are purely incentive-driven.

The AP does not know the types of competing APs. This introduces the information asymmetry: each AP knows only its own state and competitors' freeze probability γ. This way, each AP can view other APs as if they were playing mixed strategies. The freezing constraint can thus benefit the model in a technical way as well. In
citetDoraszelski2010, such modification to the original
citetPakesMcGuire model helped to prove that there exists a pure-strategy MPE and effectively eliminated the existence of mixed strategies in equilibrium.

Since the AP does not know which competitors are frozen and which are not, it will use a binomial distribution over the possible number of APs that created/redeemed today, which I denote by random variable M, where $M \sim Binominal(N, \gamma)$. If APs are identical in (λ^i, γ^i), they will have the exact same policy function and value function, conditional on not being frozen, and they will also have the same frozen value function.

The Bellman equation can be written similarly to the oligopoly case, with the distinction that now I will compute the expected value tomorrow over not only the distribution of possible ϵ but also over the number of competing APs that will engage today (M).

$$V_i(\mu) = \max_{x_i} \quad v_i(CF(x_i, \mu)) + \beta E^{M \times \epsilon}\Big[(1-\gamma)V_i(\mu') + \gamma V_i^{freeze}(\mu')\Big]$$

s.t.

$$\mu' = \rho\psi\big(x_i + X_{-i}(M)\big)u$$

$$|x_i| \geq m$$

where $X_{-i}(M)$ denotes the sum of optimal responses of competing APs that engage in creating/redeeming today, whose number is a random variable not known at time t to agent i.

The value function at freeze can be defined as follows:

$$V_i^{freeze}(\mu) = 0 + \beta E^{M \times \epsilon}\Big[(1-\gamma)V_i(\mu') + \gamma V_i^{freeze}(\mu')\Big]$$

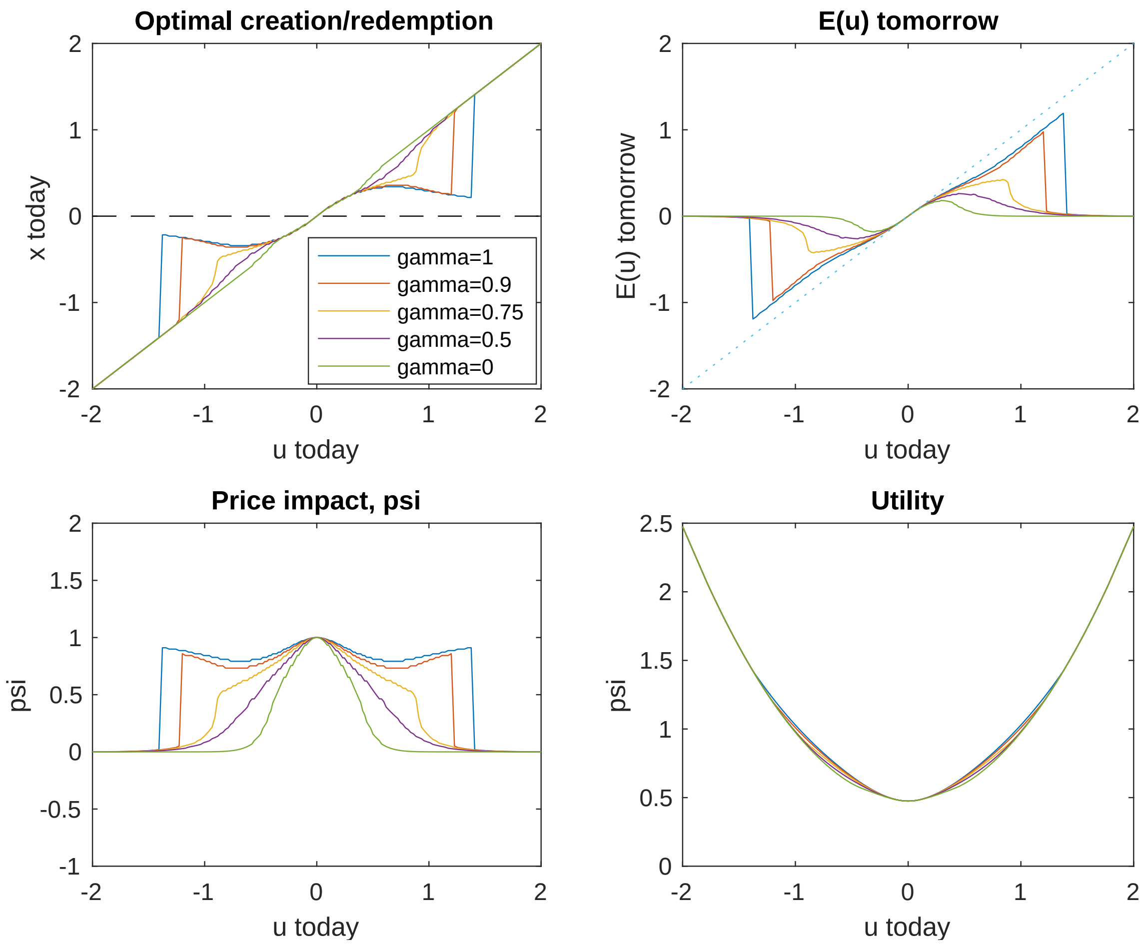
s.t.

$$\mu' = \rho\psi\big(0 + X_{-i}(M)\big)u$$

Figure 5.2 shows the model solution to the duopoly case. In case $\gamma = 1$ we have effectively the monopoly case. In case $\gamma = 0$ we have the duopoly case where AP agents always engage at the same time and at the same amounts. For $\gamma \in (0, 1)$ we can see that the model solution lies somewhere in between the two extremes.

Case 5: Model with (discrete) creation unit size

Figure 5.2: Two period model with freeze constraint



Parametrization: $\lambda = 0.5$, $\phi = 0.5$, $\rho = 0.95$, $\sigma = 1$, $\beta = 0.95$.

Source: Author's computations.

Effectively, this would be the exact same model but with grid optimization, where the grid interval represents the creation unit size. The emphasis in this case would be on the x-grid structure, which should be coarse.

5.3. Model with heterogeneous APs

In this specification, I will assume that each AP is characterized by a unique set of parameters (λ^i, γ^i). Having different levels of marginal transaction cost among APs seems to be a realistic assumption. We can think about APs that are more specialized in a particular ETF's market as those with low λ^i. APs that are less specialized in a particular ETF's market would have higher λ^i. Similarly, APs with higher γ^i can be viewed as those that transact in the market often, follow the evolution of the mispricing, and have their capital or balance sheet positions ready to transact in this primary market.

In this case, each AP will have a different cash flow function $CF_i(x_i, u)$ and its own separate active value function $V_i(u)$, a value function if nonactive (frozen) $V_i^f(u)$, and a policy function $x_i(u)$. The algorithm used to find the MPE for this model will need to be adjusted. Because introducing heterogeneity and losing the symmetry of the MPE equilibrium significantly reduces the convergence speed of the algorithm, I focus on the case with up to four APs, which corresponds to the average number of active APs. The oligopolistic structure argument for the ETF market would not be very prevalent for ETFs with a large number of active APs anyway. All the technical details and the algorithm for solving this model are included in Appendix A.5.

Another interesting consideration regarding the heterogeneous model is a differentiation between non-active and active APs. We could think about non-active APs as those with large λ^i, which prevents them from having arbitrage incentives to be active in the market given the large creation unit size and possibly, but not necessarily, large γ^i, which can capture, for example, outside investment opportunities or capital frictions. This argument, of course, would not hold for a continuous level of feasible creation redemption x.

5.4. Model calibration

The baseline model was calibrated to represent a hypothetical ETF, whose simulated moments resembles the observed moments of ETFs characterized by few active APs. The relevant data moments for ETFs with a number of active APs between 0 and 4 are reported in Table 5.1. The grid for mispricing u is defined as fairly dense with 160 grid points on interval $[-5, 5]$. The u-grid points are densest around zero and get sparser towards the two ends, as the mispricing distribution is concentrated in the center of the grid. The grid for creations/redemptions, the x-grid, allows for up to five creation units.[25]

[25] Setting it to 5 in the current model calibration does not realistically limit the APs' actions. As shown in the figures, the policy function never reaches 5 as an optimal creation/redemption for any point on the

Table 5.1: Data moments for ETFs with # active APs<=4 used to calibrate the model

# active APs			
	HHI Index	(daily) autocorrelation of mispricing	standard deviation of mispricing
0	1	0.57	0.92
1	1	0.49	0.77
2	0.65	0.44	0.63
3	0.51	0.37	0.54
4	0.45	0.39	0.47
	% of non zero fund flow days with 1 CU	% of non zero fund flow days with <=3 CU	% of days with non zero fund flows
1	63%	84%	2%
2	55%	81%	5%
3	46%	77%	8%
4	39%	71%	12%

Source: Author's computations.

The autocorrelation parameter in the low of motion for the mispricing ρ is set to 0.6, which corresponds to the daily autocorrelation detected for ETFs with zero active APs (as shown in Table A.12). The standard deviation of the demand shock σ is set to 0.75 so that the model's implied unconditional standard deviation of the mispricing u when there are no active APs (assuming u is normally distributed) equal to $\sqrt{\sigma^2/(1-\rho^2)}$ is close to the observed standard deviation of the daily mispricing for ETFs with zero active APs. Beta is set to 0.999. Marginal productivity parameters λ^i, freeze parameters γ^i, and creation unit size are set so that the HHI, fund flow frequency, and created units match the data moments. The price impact parameter ϕ is chosen to match the levels of autocorrelation of mispricings when the number of active APs is greater than zero.

Figure 5.3 shows the results from the model for baseline calibration. The model is simulated 1000 times for a period of 252 days so that one simulation corresponds to one year. In the top left panel, we can see the policy function, if non-frozen, for each of the heterogeneous APs, which are ordered from most to least specialized in this market as measured by their marginal transaction cost. The policy function has a stair-step shape showing exactly different thresholds in the level of mispricing for the optimality of one or two creation units

u-grid.

among APs. The bottom panel shows the distribution of mispricings in the model when there are no APs and how this distribution looks when four "potential" APs are present. Table 5.2 reports some basic moments for simulated data in the baseline calibration that can be easily compared with the real data moments reported in Table 5.1. We can see that the average number of active APs in the simulation was 2.4, the average HHI was 0.67, and the average autocorrelation of mispricing was 0.41. All those moments are not far from the real data moments for ETFs with 2 or 3 APs.

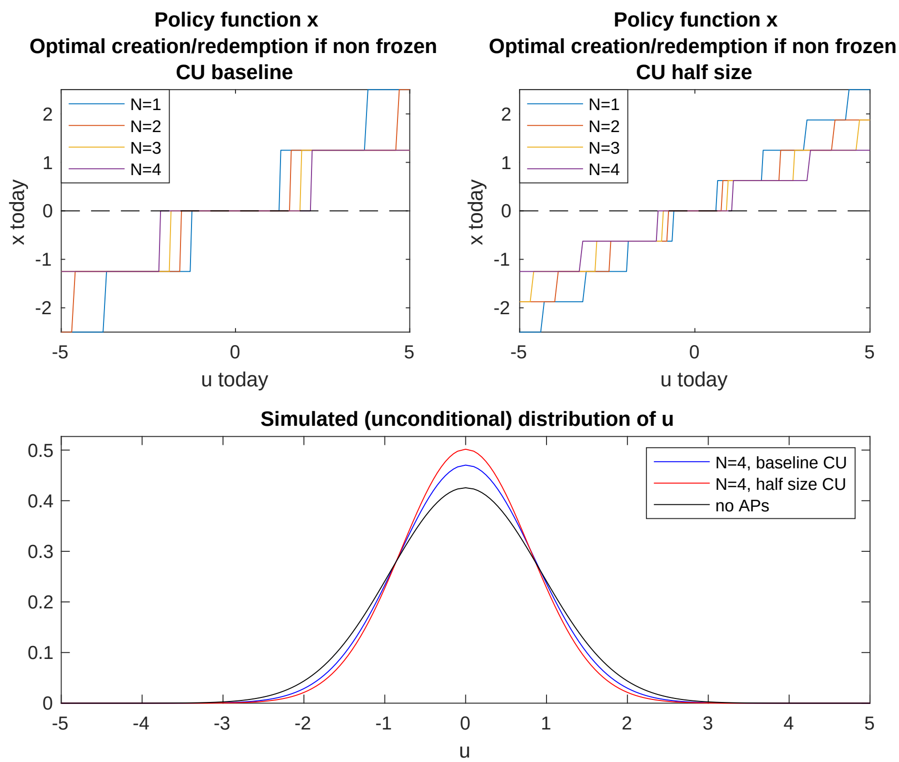
Another important aspect is that the data moments are computed for all ETFs with a certain number of authorized participants. However, not all of them need to suffer from a creation unit size that poses a friction. In a particular ETF, even with a relatively small number of active APs, which shares are created and redeemed frequently and in many multiples of the creation unit, clearly the creation unit size is not a relevant friction. It is the ETFs that show very rare non-zero fund flow days, with the majority of them at one created unit, that are more likely to be subject to this friction.

5.5. Model implications for creation unit size change

After calibrating the model, I perform policy analysis for a different level of creation unit size to evaluate its effect on the mispricing levels. Top right panel of figure 5.3 shows the model policy function. We can see that the APs are starting to transact in the primary market at the lower levels of mispricing. Also, the distribution of mispricing is slimmer than in the baseline case. Simulated data moments are also reported in Table 5.2. The policy experiment of lowering the creation unit size by half increased the average number of active APs from 2.4 to 3.8, decreased HHI index from 0.67 to 0.56, and decreased the standard deviation of mispricing by 6 basis points. This corresponded to a decrease in the average absolute mispricing of 5 basis points. This is slightly higher than the result from the reduced form analysis, where a similar effect was estimated to lower the absolute mispricing by around 2 basis points. However, it is worth pointing out that the reduced form regression covered events across the whole space of ETFs, including ETFs with more than four active

APs; thus, the average level of absolute mispricing in the reduced form regressions was smaller to begin with. The frequency of fund flows increased from 13% to 43%.

Figure 5.3: Heterogenous model with freeze - policy function and distribution of mispricing



Source: Author's computations.

Although the model does not necessarily answer the question of the optimality of the creation unit size, it still offers a very useful framework for the ETFs that are considering adjusting their creation unit size. The reduced form estimates are rather unreliable to guide such decisions given the small number of such events in the whole cross-section of ETFs, let alone in the individual ETF history itself, or even in the group of similar ETFs. The proposed structural framework, when calibrated to match the individual ETF primary market characteristics and the characteristics of their APs, can provide a useful workhorse for evaluating what the response would be to different creation unit size policies from the APs' side and how would it affect the level of mispricing. This would be a relevant ingredient in the design of the optimal size of the creation unit size from the perspective of the ETF issuer, who in the end is in charge of its level. Such an analysis would, however, require a much broader discussion on the ETF issuer's costs and benefits of having a larger versus smaller creation unit size due to, for example, the trade-off between smaller mispricings potentially attracting more capital, difficulties in replicating tracked indices, the need for higher cash positions, or simply the higher administrative cost of managing more frequent creation and redemption orders.

Table 5.2: Model simulated moments for model with N=4

	Baseline	**0.8 CU**	**Half CU**
CU size	1.25	1	0.625
Calibration			
λ_i		{1, 1.25, 1.5, 1.75}	
ϕ		5	
σ		0.75	
ρ		0.6	
β		0.999	
γ_i		{0.05 0.5 0.95 0.95}	
Average data moments across 1000 simulations			
standard deviation of mispricings	0.85	0.82	0.79
average of absolute mispricings	0.68	0.66	0.63
autocorrelation of mispricings	0.41	0.34	0.28
# active AP	2.39	2.85	3.83
HHI index	67%	63%	56%
% transaction days with 1 CU	67%	43%	23%
% of transaction days with <=3 CU	93%	85%	59%
% of days with non zero fund flows	13%	21%	43%

Note: 1000 simulations, each 252 periods long.

Source: Author's computations.

CHAPTER 6

Conclusions

This paper documents new stylized facts about the primary market of exchange traded funds (ETFs) in the United States. First, it shows that a large portion of US listed ETFs are characterized an oligopolistic structure. Three out of ten US ETFs have no more than two APs actively transacting in their primary market. Also, the average of the HHI concentration index for the whole universe of US listed ETFs is 0.5. Together with the average number of transacting APs per ETF, this suggests that, on average, the market shares between APs are asymmetric, typically with one bigger and two or three smaller APs.

Second, the primary market structure of the ETF is empirically shown to have important implications for ETF pricing efficiency. A small number of APs and a high level of primary market concentration are associated with higher magnitudes of mispricings.

Third, the paper also investigates the characteristics of the creation unit size. Since the cases of organic changes in creation unit sizes are rather rare, ETF splits are also considered when estimating the effect of a shock to the creation unit size. The results suggest that, indeed, ETF market efficiency improves in terms of lower ETF mispricings in the aftermath of such shocks to the dollar value of the creation unit size.

Lastly, this paper proposes a framework that can be used to model the intertemporal choice that an AP faces as an arbitrageur and the effect of competition on the evolution of the mispricing in the ETF market. The model calibrated to match the moments of the ETFs with the least competitive primary market provided an interesting policy experiment result regarding the creation unit size. A decrease in creation unit size by half could result in a 6 basis point decrease in the standard deviation of the mispricing for ETFs with an oligopolistic primary market structure and would result in one extra active AP on average in those ETFs.

APPENDIX A

Extra Tables and Figures

Table A.1: Description of Relevant Variables used in the Empirical Analysis

Name	Definition	Source
apN	Number of registered APs (number of APs listed in the N-CEN filing for a particular ETF)	EDGAR
apNactive	Number of registered APs who created or redeemed ETF shares for a particular ETF during the N-CEN filing period	EDGAR
apActiveFreq	Percent of active APs = apNactive / apN	EDGAR
HHI creation	HHI index for creations, $HHI = \sum_j \left(\frac{S^j}{\sum_j S^j} \right)^2$ where S^i_j is an absolute dollar value of creations of AP j	EDGAR
HHI redemption	HHI index for redemptions	EDGAR
HHI volume	HHI index for volume (volume defined as a sum of absolute creation and absolute redemption)	EDGAR
mispricing	Difference between ETF price and NAV at the end of the day, absolute mispricing in $, relative mispricing in p.p.	CRSP, Bloomberg, ETFGlobal
creation unit	Creation Unit Size	ETFGlobal, EDGAR
creation fee	Creation FEE	ETFGlobal, EDGAR
age	years since inception date	ETFGlobal
aum	Assets Under Management	ETFGlobal, CRSP, Bloomberg
asset class	ETF asset class category	ETFGlobal
fund flows	ETF daily fund flows	ETFGlobal, Bloomberg
volume	ETF daily volume	CRSP
bid ask spread	ETF bid ask spread *100*(ask-bid)/ask*	CRSP
constituents bid ask spread	wieghted daily bid ask spread of ETF constituents, using constituents weights from ETFGlobal, computed only for equity ETFs for which at least 95% constituents as measure by weight were matched successfully with CRSP database	CRSP, ETFGlobal
VIX	Market Daily Volatility Index	CRSP

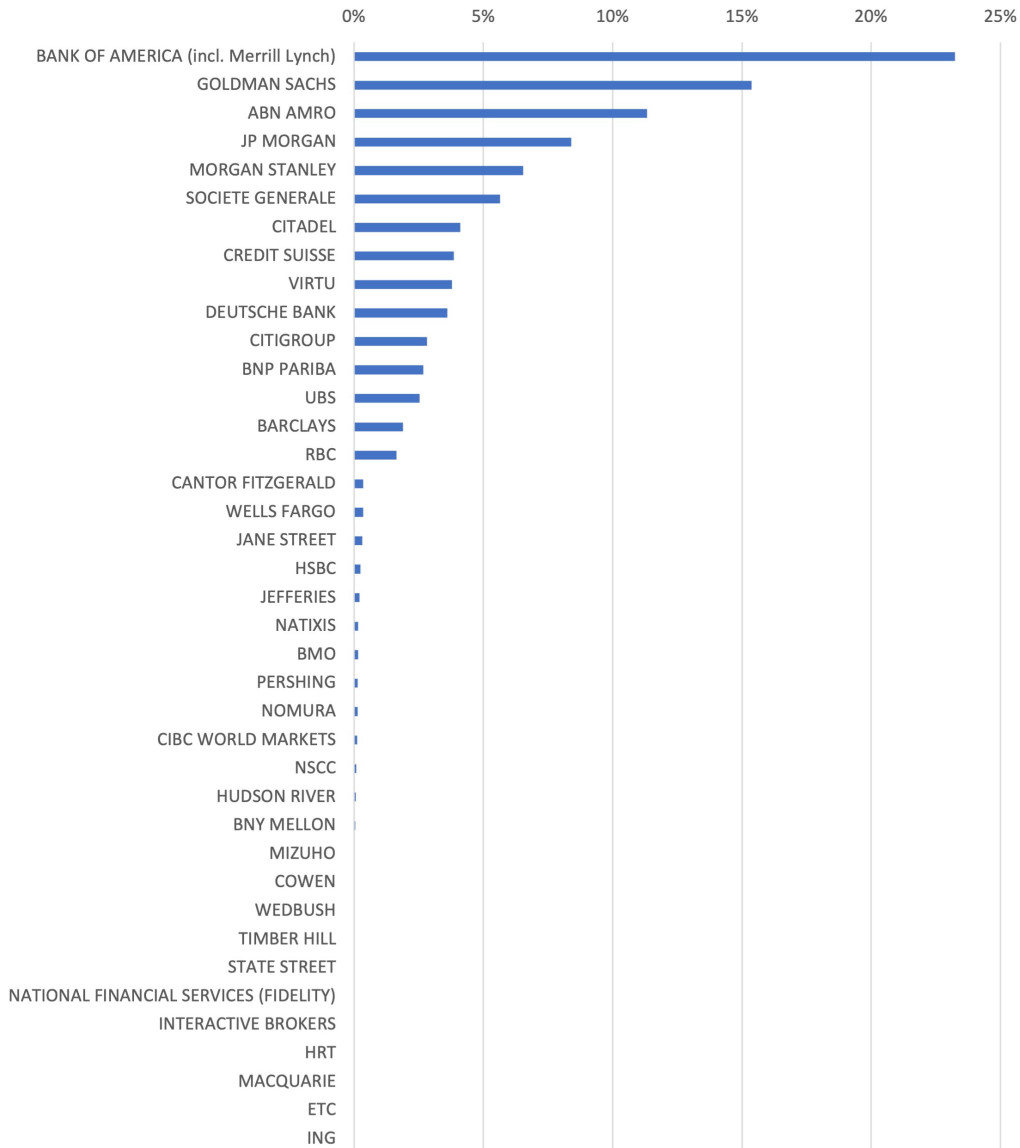
Table A.2: Authorized participants

This table presents the financial institution that appear in the N-CEN filings as authorized participants for filings sumbitted between July 2018 - August 2021. The parsed database contained over 300 different AP names with 90 unique LEI identifiers. After grouping APs at the parent organization, I was able to identify 58 unique AP names listed in the table below. Among those, only 42 had any primary market transactions in created database. Multiple LEI identifiers per group holding company are often the effect of mergers and acquisitions. For example Virtu acquired Knigth Capital in 2017 and Investment Technology Group (ITG) in 2019. Convergex got acquired by Cowen in 2017. NatWest became part of The Royal Bank of Scotland (RBS) Group in 2000 and in 2020 RBS Group was renamed NatWest Group. BMO Nesbitt Burns and BMO Capital are both part of Bank of Montreal. Newedge got acquired by Societe Generale (SG) in 2014. First Southwest Company is owned by Hilltop Secirities. Merrill Lynch was acquired by Bank of America in 2008. TD Ameritrade got acquired by Charles Schwab in October 2020 and E*Trade got acquired by Morgan Stanley in the same month.

Authorized Participants	
ABN AMRO	ITG*
BANCA IMI*	JANE STREET
BANK OF MONTREAL	JEFFERIES
BARCLAYS	JP MORGAN
BNP PARIBA	MACQUARIE
BNY MELLON	MITSUBISHI*
BANK OF AMERICA (incl. MERRILL LYNCH)	MIZUHO
CANTOR FITZGERALD	MORGAN STANLEY
CETERA*	NATIONAL BANK OF CANADA*
CIBC WORLD MARKETS	NATIONAL FINANCIAL SERVICES (FIDELITY)
CITADEL	
CITIGROUP	NATIXIS
CLEAR STREET	NEWEDGE*
COMMERZBANK*	NOMURA
COWEN	NSCC
CREDIT SUISSE	PERSHING
DAIWA*	RBC
DEUTSCHE	RWBAIRD
ETC	SCOTIA*
FIRST SOUTHWEST*	SOCIETE GENERALE
FLOW TRADERS*	STATE STREET
GOLDMAN	STIFEL NICOLAUS*
HILLTOP*	TD
HSBC	TIMBER HILL
HUDSON RIVER	UBS
INDUSTRIAL AND COMERCIAL BANK OF CHINA*	VIRTU
ING	VIRTUS
INTERACTIVE BROKERS	WEDBUSH
ITAU*	WELLS FARGO

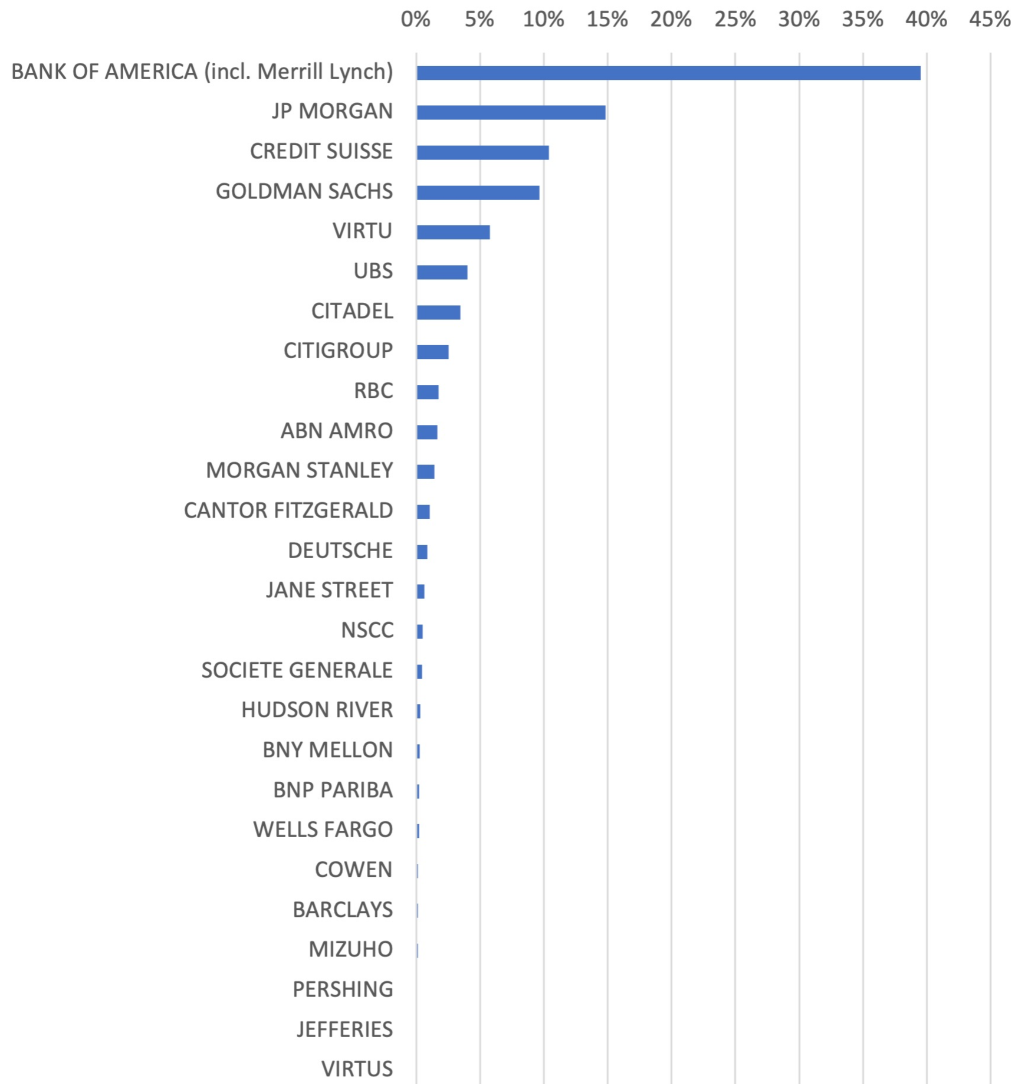
*APs with no primary market activity for collected N-CEN filings.

Figure A.1: Market share of active APs across all ETFs



Source: Own computations using N-CEN filings filed between July 2018-July 2020. Market share computed to total transacted volume in the collected database.

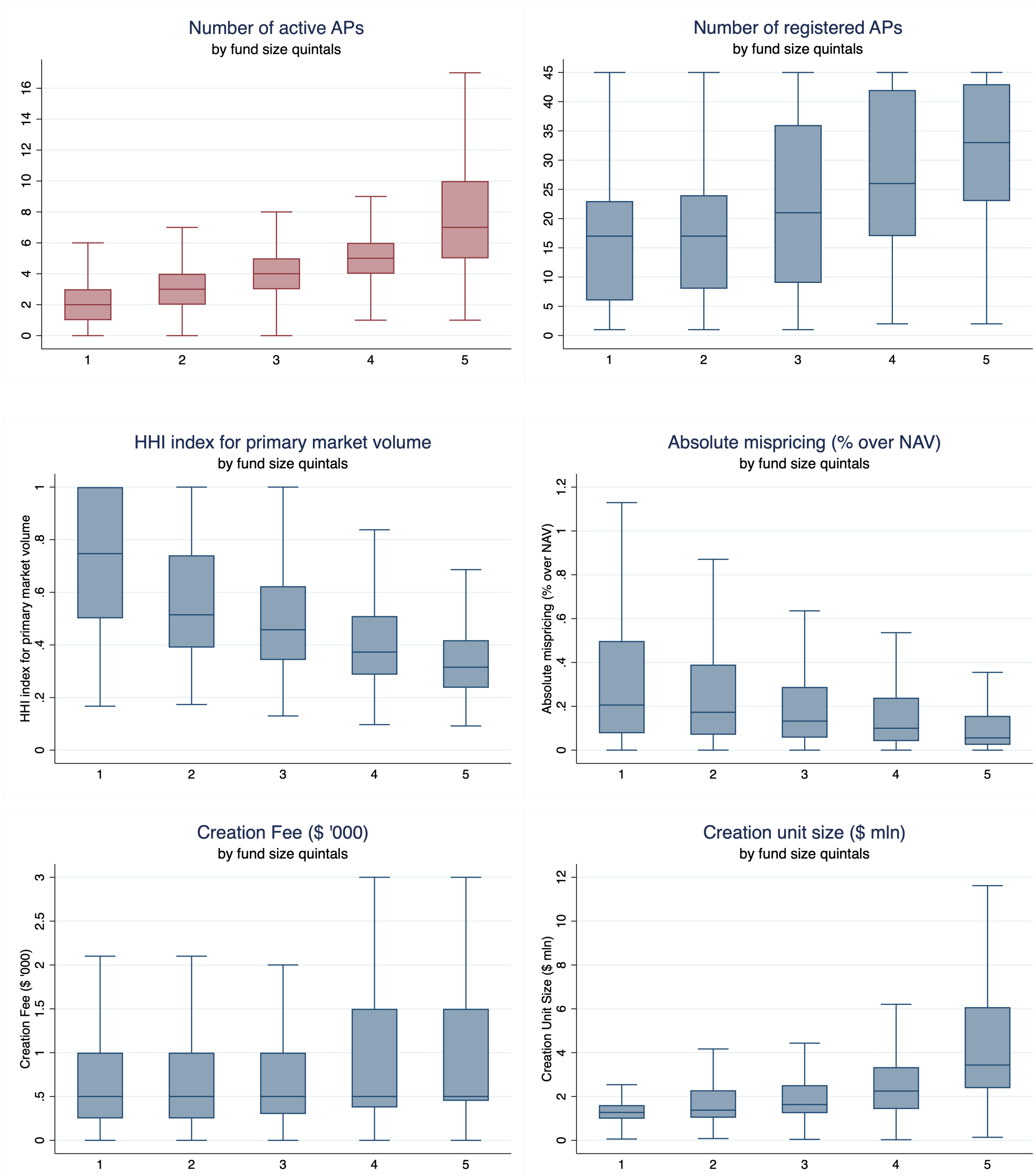
Figure A.2: Who are the biggest APs



Source: Own computations using N-CEN filings filed between July 2018-August 2021. Distribution of biggest APs per ETF filing. One biggest (active) AP as measured by total primary trasaction volume per one ETF filing selected.

Figure A.3: Data moments by fund size

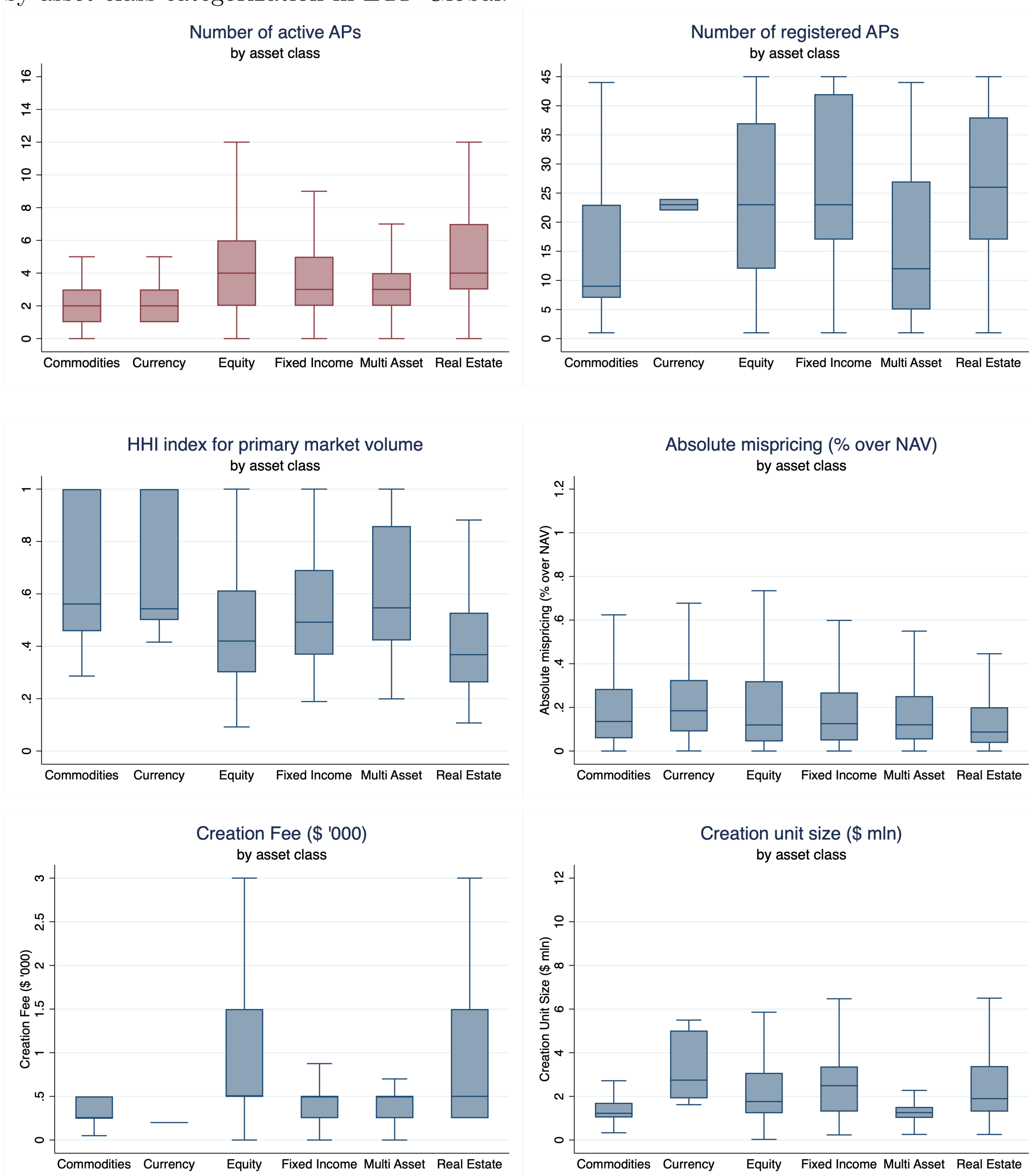
Graphs below shows the average characteristics across 5 subgroups of ETFs split by size as defined by the asset under management. Data spans the ETFs that filed N-CEN filings between July 2018 - August 2021.



Source: Author's computations.

Figure A.4: Data moments by asset class

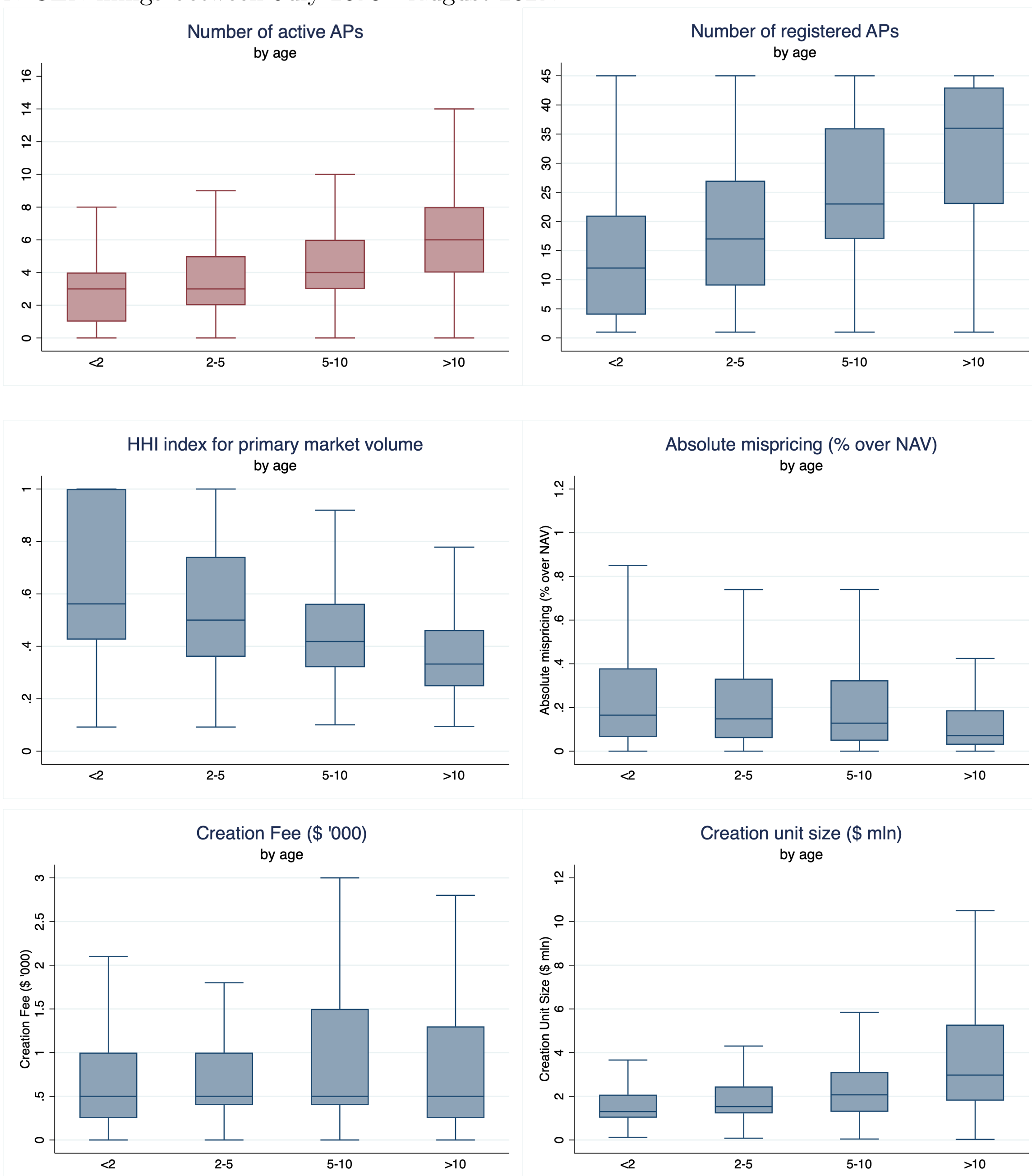
Graphs below shows the average characteristics across 6 categories of ETF asset class that filed N-CEN filings between July 2018 - August 2021. Data spans 1842 Equity, 470 Fixed Income, 123 Multi Asset, 47 Real Estate, 17 Commodities and 4 Currency ETFs, as defined by asset class categorization in ETF Global.



Source: Author's computations.

Figure A.5: Data moments by age

Graphs below shows the average characteristics across 4 age categories of ETFs that filed N-CEN filings between July 2018 - August 2021.



Source: Author's computations.

Table A.3: In-Kind portion of creation and redemption baskets by ETF class size, age and size

This table presents the median value of in-kind portion of the creation and redemption baskets. The observation unit is N-CEN filing. ETFs with zero primary market trasnactions during N-CEN report period excluded. Data span N-CEN filings filed July 2018 - August 2021.

Median In-Kind portion of creation and redemption baskets by ETF asset category, age (in columns) and size (in rows)

Creation Baskets

	Equity ETFs						*Fixed Income ETFs*				
	age (years)						**age (years)**				
size ($)	*<2*	*2-5*	*5-10*	*>10*	Total	**size ($)**	*<2*	*2-5*	*5-10*	*>10*	Total
<24mln	97	99	91	0	97	*<24mln*	0	91	0	0	0
24-110mln	96	98	98	100	98	*24-110mln*	0	0	50	0	0
110-650mln	97	99	97	99	99	*110-650mln*	3	29	38	43	35
>650mln	99	99	97	99	98	*>650mln*	44	63	75	95	87
Total	97	99	97	99	98	Total	0	17	56	87	38

Redemption Basekts

	Equity ETFs						*Fixed Income ETFs*				
	<2	*2-5*	*5-10*	*>10*	Total		*<2*	*2-5*	*5-10*	*>10*	Total
<24mln	98	99	96	0	97	*<24mln*	94	70	0	0	50
24-110mln	97	99	98	99	98	*24-110mln*	67	39	81	0	51
110-650mln	99	99	98	99	99	*110-650mln*	0	67	86	88	76
>650mln	99	99	98	99	99	*>650mln*	48	48	93	98	96
Total	98	99	98	99	99	Total	78	54	87	98	86

Count of N-CEN reports with primary market volume>0

	Equity ETFs						*Fixed Income ETFs*				
	<2	*2-5*	*5-10*	*>10*	Total		*<2*	*2-5*	*5-10*	*>10*	Total
<24mln	520	276	131	95	1022	*<24mln*	118	30	11	6	165
24-110mln	257	315	237	159	968	*24-110mln*	101	88	44	15	248
110-650mln	77	255	295	365	992	*110-650mln*	31	107	109	49	296
>650mln	21	81	282	656	1040	*>650mln*	6	35	138	134	313
Total	875	927	945	1275	4022	Total	256	260	302	204	1022

Source: Author's computations using N-CEN filings.

Table A.4: Type of fees charged by ETFs for creation/redemptions, charge style

This table presents the three-way distribution of types of creation fees charged by the ETFs. The observation unit is N-CEN filing. Empty and zero answers treated as no fees charged. Data span N-CEN filings filed August 2018 - August 2021.

A: The average transaction fee charged to an authorized participant for transacting in the creation units, expressed as:

	Charge per dollar value			
	No		*Yes*	
	Charge per creation unit		Charge per creation unit	
Charge per order	*No*	*Yes*	*No*	*Yes*
No	906	354	44	57
Yes	4305	19	485	21

B: The average transaction fee charged to an authorized participant for transacting in those creation units the consideration for which was fully or partially composed of cash, expressed as:

	Charge per dollar value			
	No		*Yes*	
	Charge per creation unit		Charge per creation unit	
Charge per order	*No*	*Yes*	*No*	*Yes*
No	1218	220	64	55
Yes	4115	20	478	21

C: The average transaction fee charged to an authorized participant for transacting in the redemption units, expressed as:

	Charge per dollar value			
	No		*Yes*	
	Charge per creation unit		Charge per creation unit	
Charge per order	*No*	*Yes*	*No*	*Yes*
No	1238	277	22	35
Yes	4275	22	318	4

D: The average transaction fee charged to an authorized participant for transacting in those redemption units the consideration for which was fully or partially composed of cash, expressed as:

	Charge per dollar value			
	No		*Yes*	
	Charge per creation unit		Charge per creation unit	
Charge per order	*No*	*Yes*	*No*	*Yes*
No	1631	246	13	35
Yes	3926	21	316	3

Source: EDGAR, N-CEN filings.

Table A.5: Fees for creation units charged by ETFs, summary statistics

This table presents the summary statistics for creation fees charged by the ETFs. The observation unit is N-CEN filing. Only non-zero answers included. Data span N-CEN filings filed July 2018 - August 2021.

	N	P1	P5	P25	Median	P75	P95	P99
Creations fees								
The average transaction fee charged to an authorized participant for transacting in the creation units, expressed as:								
Dollars per creation unit, if charged on that basis	451	1	44	125	250	500	3,586	11,323
Dollars for one or more creation units purchased on the same day, if charged on that basis:	4,830	0	150	250	500	1,150	6,000	147,000
A percentage of the value of each creation unit, if charged on that basis: [%]	564	0.00	0.00	0.01	0.05	0.16	2.00	3.81
The average transaction fee charged to an authorized participant for transacting in those creation units the consideration for which was fully or partially composed of cash, expressed as:								
Dollars per creation unit, if charged on that basis	316	0	50	222	348	934	6,222	91,626
Dollars for one or more creation units purchased on the same day, if charged on that basis:	4,634	0	150	250	500	1,148	6,100	147,000
A percentage of the value of each creation unit, if charged on that basis: [%]	591	0.00	0.00	0.03	0.09	0.30	2.00	3.00
Redemptions fees								
The average transaction fee charged to an authorized participant for transacting in the redemption units, expressed as:								
Dollars per creation unit, if charged on that basis	338	10	42	205	391	712	7,000	22,882
Dollars for one or more creation units purchased on the same day, if charged on that basis:	4,619	20	150	250	500	1,000	5,100	133,000
A percentage of the value of each creation unit, if charged on that basis: [%]	362	0.00	0.00	0.01	0.05	0.25	2.00	3.76
The average transaction fee charged to an authorized participant for transacting in those redemption units the consideration for which was fully or partially composed of cash, expressed as:								
Dollars per creation unit, if charged on that basis	305	13	75	250	500	960	7,000	22,882
Dollars for one or more creation units purchased on the same day, if charged on that basis:	4,266	31	158	250	500	1,100	5,500	133,000
A percentage of the value of each creation unit, if charged on that basis: [%]	362	0.00	0.00	0.05	0.15	0.44	2.00	2

Source: EDGAR, N-CEN filings

Table A.6: Summary statistics, sub sample used in regressions of non-levered ETFs with HHI>0.33

This table presents summary statistics for basic variables used in regressions. Data span N-CEN filings filed between July 2018 - August 2021. Observations are daily. Weighted Bid Ask Spread is computed for ETFs for which constituents with total weight of at least 95% were successfully matched with CRSP stock database.

	N	Mean	Sd	P1	P25	Median	P75	P99
Mispricing								
P_t - NAV_t ($)	754,521	0.01	0.16	-0.64	-0.04	0.01	0.06	0.54
100*(P_t-NAV_t)/NAV_t	754,521	0.01	0.45	-1.72	-0.12	0.02	0.18	1.45
Absolute mispricing								
abs(P_t - NAV_t)	754,521	0.10	0.14	0.00	0.02	0.05	0.12	0.81
100*abs(P_t-NAV_t)/NAV_t	754,521	0.29	0.39	0.00	0.06	0.15	0.36	2.20
Age (years)	754,429	6.2	4.3	1.1	2.7	5.0	9.0	19.3
AUM ($ bln)	754,053	1.1	3.8	0.0	0.0	0.1	0.4	22.1
Creation Unit ($ mln)	753,465	2.6	2.2	0.5	1.2	1.8	2.9	13.1
Creation Fee (bp of 1 CU)	753,465	0.8	1.1	0.0	0.2	0.4	0.9	6.0
Volume ($ '000)	754,520	13	68.2	0.0	0.1	0.4	2.5	302.0
Bid Ask Spread (% over Ask)	754,312	0.28	0.32	0.01	0.08	0.17	0.35	1.81
Weighted Bid Ask Spread	238,157	0.11	0.12	0.01	0.03	0.06	0.13	0.63
VIX	754,521	19.9	9.8	9.6	13.3	16.3	23.2	64.0

Source: Author's computations.

Table A.7: Effect of primary market structure on ETF price deviations from NAV, fund and time fixed effects

This table presents the results of pooled OLS regressions of absolute mispricing relative to NAV on market structure including number of registered APs, active APs and HHI index computed using primary market transactions volumes on the absolute mispricing. Sample includes ETFs older than 1 year and with HHI index above 0.33. Levered ETFs are excluded. Dates for which the reporting period of the N-CEN was less than a year are excluded. Regressions with constituents bid ask spread, columns (4)-(6), contains ETFs for which at least 95% constituents (as measured by weight) were matched with CRSP stock data. Observations are daily. Number of registered APs, active APs and HHI index is from the ETF's N-CEN filing with reporting period that overlapped with the daily observation. Data span the reporting periods for the N-CEN filings filed by the ETF between July 2018 - August 2021, thus the starting dates included in the sample are July 2017 for ETFs with fiscal year ending in June and later for fiscal year ending in other months.

	Relative Asbolute Mispricing (over NAV, in basis points)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of registered APs	-0.07			-0.03	-0.2*			-0.2
	(0.06)			(0.06)	(0.1)			(0.1)
Number of active APs		-0.8***		-0.7**		-1.0**		-0.5
		(0.3)		(0.3)		(0.5)		(0.5)
HHI of primary market (volume)			3.2*	0.9			6.6**	4.7
			(1.7)	(1.8)			(3.2)	(3.2)
Creation Unit [mln $ previous month]	0.08	0.07	0.1	0.06	-0.7	-0.6	-0.5	-0.8
	(0.4)	(0.4)	(0.4)	(0.4)	(0.7)	(0.7)	(0.7)	(0.7)
Creation fee [bp of 1CU previous month]	1.4**	1.4**	1.4**	1.4**	3.1**	2.8**	3.0**	3.0**
	(0.6)	(0.6)	(0.6)	(0.6)	(1.2)	(1.2)	(1.2)	(1.2)
AUM [bln $ previous month]	0.1	0.08	0.10	0.09	1.0**	1.0*	1.0**	1.0*
	(0.1)	(0.1)	(0.1)	(0.1)	(0.5)	(0.5)	(0.5)	(0.5)
Volume [mln $ previous month]	0.003	0.003	0.003	0.003	0.02**	0.02**	0.02***	0.02**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.007)	(0.007)	(0.007)	(0.007)
Bid Ask Spread [% previous day]	11.7***	11.7***	11.7***	11.7***	9.6***	9.5***	9.6***	9.5***
	(1.3)	(1.3)	(1.3)	(1.3)	(1.7)	(1.7)	(1.7)	(1.7)
Vix	1.0***	1.0***	1.0***	1.0***	0.9***	0.9***	0.9***	0.9***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
Constituents Bid Ask Spread [% previous day]					6.8***	6.9***	7.0***	6.9***
					(2.2)	(2.2)	(2.3)	(2.2)
N	754293	754293	754293	754293	237963	237963	237963	237963
Num of ETFs	1693	1693	1693	1693	651	651	651	651
R squared	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
ETF fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
YearMonth fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SE clustered at the ETF level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses * $p<0.1$ ** $p<0.05$ *** $p<0.01$

Source: Author's computations.

Table A.8: Effect of primary market structure on ETF price deviations from NAV, cross variation

This table presents the results of pooled OLS regressions of absolute mispricing relative to NAV on market structure including number of registered APs, active APs and HHI index computed using primary market transactions volumes on the absolute mispricing. Sample includes ETFs older than 1 year and with HHI index above 0.33. Levered ETFs are excluded. Dates for which the reporting period of the N-CEN was less than a year are excluded. Regressions with constituents bid ask spread, columns (4)-(6), contains ETFs for which at least 95% constituents (as measured by weight) were matched with CRSP stock data. Observations are daily. Number of registered APs, active APs and HHI index is from the ETF's N-CEN filing with reporting period that overlapped with the daily observation. Data span the reporting periods for the N-CEN filings filed by the ETF between July 2018 - August 2021, thus the starting dates included in the sample are July 2017 for ETFs with fiscal year ending in June and later for fiscal year ending in other months.

	Relative Asbolute Mispricing (over NAV, in basis points)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of registered APs	-0.02			0.07*	-0.09*			0.02
	(0.04)			(0.04)	(0.05)			(0.05)
Number of active APs		-2.7***		-2.4***		-3.1***		-2.9***
		(0.2)		(0.2)		(0.5)		(0.4)
HHI of primary market (volume)			17.2***	4.6**			17.9***	3.4
			(2.1)	(2.3)			(4.1)	(4.0)
Creation Unit [mln $ previous month]	1.0***	0.9***	0.9***	0.8***	0.2	0.007	-0.002	-0.04
	(0.3)	(0.2)	(0.2)	(0.2)	(0.3)	(0.3)	(0.3)	(0.3)
Creation fee [bp of 1CU previous month]	6.3***	5.7***	6.1***	5.7***	6.4***	5.7***	6.1***	5.7***
	(0.8)	(0.7)	(0.8)	(0.7)	(0.9)	(1.0)	(0.9)	(1.0)
AUM [bln $ previous month]	-0.5***	-0.2*	-0.4***	-0.2*	-1.0***	0.2	-0.8**	0.1
	(0.1)	(0.09)	(0.10)	(0.09)	(0.3)	(0.3)	(0.3)	(0.3)
Volume [mln $ previous month]	0.01***	0.01***	0.01***	0.01***	0.03**	0.04***	0.04***	0.04***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.01)	(0.01)	(0.01)	(0.01)
Bid Ask Spread [% previous day]	24.2***	21.4***	22.8***	21.5***	25.1***	22.6***	24.2***	22.7***
	(2.0)	(2.0)	(2.0)	(2.0)	(3.2)	(3.2)	(3.2)	(3.2)
Vix	0.9***	0.9***	0.9***	0.9***	0.6***	0.6***	0.6***	0.6***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Constituents Bid Ask Spread [% previous day]					20.9***	20.0***	21.8***	20.3***
					(4.7)	(4.7)	(4.7)	(4.7)
N	754238	754238	754238	754238	237959	237959	237959	237959
Num of ETFs	1691	1691	1691	1691	651	651	651	651
R squared	0.15	0.16	0.15	0.16	0.15	0.16	0.16	0.16
ETF Asset Class Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SE clustered at the ETF level	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses * $p<0.1$ ** $p<0.05$ *** $p<0.01$

Source: Author's computations.

Table A.9: ETF Pairs Analysis - regressions with ETF pair fixed effect

This table reports the results of the regressions for ETF pairs. ETF pairs were constructed as two ETFs with the same benchmark index. Unit of observation is the difference or log-difference of characteristics between two ETFs within one ETF pair. Only non-levered ETFs are considered. Standard deviations and the frequency of non-zero fund flow days are computed over each quarter using daily observations. Regressors are computed as median observation each quarter. Data spans 2017 Q3 - 2021 Q2.

	Diff in StD of Relative Mispricing (in basis points)		Diff in StD of fund flow (in basis points)		Diff in frequency of days with nonzero fundflows (in %)	
	(1)	(2)	(3)	(4)	(5)	(6)
Diff in \# active APs	-0.4*		6.7		0.8*	
	(0.2)		(4.7)		(0.4)	
Diff in HHI_volume		-1.8		-63.7*		-11.4***
		(6.4)		(32.6)		(4.2)
ln(Ratio of Creation Unit Size in Dollars)	0.5	1.0	-9.6	-12.7	-6.4*	-6.5*
	(1.5)	(1.3)	(41.2)	(40.3)	(3.5)	(3.5)
ln(Ratio of AUM)	-3.4	-3.6	12.3	15.8	8.6***	8.8***
	(2.3)	(2.4)	(46.4)	(47.0)	(1.6)	(1.5)
Observations	759	759	697	697	740	740
r2	0.6	0.6	0.3	0.3	0.9	0.9
SE clustered at	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair
Fixed Effect	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair	ETFpair
/# of ETF pairs	85	85	82	82	83	83

Standard errors in parentheses

="* p<0.1 ** p<0.05 *** p<0.01"

Source: Author's computations.

Table A.10: Asymetric effect of the split on the mispricing magnitudes

This table shows the effect of ETF splits (not accompanied by the simultaneous decrease in creation unit size) and of the decrease in Creation Unit Size (not accompanied by the ETF split) on the bid ask spreads. Creation Unit Size change events are identified using daily data in creation unit size from ETF Global, which is available starting from year 2012. ETF split events are identified using the cumulative adjustment factor for shares (cfacshr) in CRSP. Magnitude variable represents the magnitude of the split or decrease and is larger than 1. Daily observations around the event are included from 1 to 90 days prior and post event. Leveraged and volatility focused ETFs are excluded. Variables dummyDecrease and dummySplit variables equal 0 prior to the event and 1 after the event. Variables premiumDummy equals 0 for discounts and 1 for premiums. Levered, volatility and long-short ETFs are excluded. Mispricing computed at the end of day.

	Absolute Relative Mispricing (over NAV, in bp)	
	(1)	(2)
premiumDummy=0 # dummySplit # ln(Magnitude)	-4.05**	
	(1.80)	
premiumDummy=1 # dummySplit # ln(Magnitude)	-1.30	
	(0.97)	
premiumDummy=0 # dummyDecrease # ln(Magnitude)		-1.72
		(1.27)
premiumDummy=1 # dummyDecrease # ln(Magnitude)		-3.20**
		(1.58)
premiumDummy=1	-0.32	5.53***
	(1.67)	(1.86)
Observations	13505	3852
R-squared	0.569	0.373
Fixed Effect	Event	Event
SE clustered at event level	Event	Event x YearMonth
# of events	158	39
Events starting date	2011	2016

Standard errors in parentheses

* p<0.1 ** p<0.05 *** p<0.01

Source: Author's computations.

Table A.11: Effect of ETF share split or decrease in creation unit size on bid ask spread

This tables shows the effect of ETF splits (not accompanied by the simultaneous decrease in creation unit size) and of the decrease in Creation Unit Size (not accompanied by the ETF split) on the bid ask spreads. Creation Unit Size change events are identified using daily data in creation unit size from ETF Global, which is available starting from year 2016. ETF split events are identified using the cumulative adjustment factor for shares (cfacshr) in CRSP. Magnitude variable represents the magnitude of the split or decrease and is larger than 1. Daily observations around the event are included from 1 to 15 days prior and post event. Leveraged and volatility focused ETFs are excluded. Only ETFs older than 1 year are included. Dummy variable equals 0 prior to the event and 1 after the event. Bid Ask Spread computed at the end of day.

	Bid Ask Spread (over Ask, in bp)	
	(1)	(2)
dummyDecrease # ln(Magnitude)	3.4*	
	(1.7)	
dummySplit # ln(Magnitude)		4.2**
		(1.7)
Observations	644	2276
R-squared	0.722	0.679
Fixed Effect	Event	Event
SE clustered at event level	Event x YearMonth	Event
# of events	39	158
Events starting date	2016	2011

Standard errors in parentheses

* $p<0.1$ ** $p<0.05$ *** $p<0.01$

Source: Author's computations.

Table A.12: Autocorrelation of the daily mispricing

	(1) Mispricing	(2) Mispricing
L.Mispricing	0.42***	
	(0.019)	
Num of active APs=0 × L.Mispricing		0.57***
		(0.041)
Num of active APs=1 × L.Mispricing		0.49***
		(0.036)
Num of active APs=2 × L.Mispricing		0.44***
		(0.033)
Num of active APs=3 × L.Mispricing		0.37***
		(0.045)
Num of active APs=4 × L.Mispricing		0.39***
		(0.056)
Num of active APs=5 × L.Mispricing		0.25***
		(0.027)
Num of active APs=6 × L.Mispricing		0.20***
		(0.044)
...	...	...
Num of active APs=24 × L.Mispricing		-0.0041
		(0.035)
Constant	0.016***	0.017***
	(0.00053)	(0.00050)
Observations	878846	878846
R^2	0.180	0.192
Number of ETFs	2179	2179
SE clustered at ETF level	Yes	Yes
Ticker Fixed Effect	Yes	Yes

Standard errors in parentheses

* $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$

Source: Author's computations.

APPENDIX B

Technical Appendix

A.1. Bellman Equation Monopoly

We will consider a case with bounded set of possible mispricings levels $u \in [\underline{u}, \overline{u}]$. This will ensure that the Cash Flow function $F()$ is bounded.

Recall that the infinite horizon problem can be written as the function of state variables: μ which captures the expected value of mispricing and is predetermined and shock ϵ which is realized at the beginning of the period.

$$TV(\mu, \epsilon) = \max_{\mu' \in \Gamma(\mu, \epsilon)} \quad F(\mu, \mu', \epsilon) + \beta \int V(\mu', \epsilon') dQ\epsilon'$$

or assuming $u = \mu + \epsilon$

$$TV(u) = \max_{x \in \Gamma(u)} \quad F(u, x) + \beta \int V(u') dQ\epsilon'$$

Assumptions:

(i) $u \in U$, U is a convex set

(ii) $F()$ is continuous and bounded

(iii) $\beta \in [0, 1)$

(iv) $\Gamma()$ is non-empty, continuous and compact-valued

(v) Q possesses the Feller property

Under assumptions (i)-(v) the Bellman operator T has a unique fixed point V in the space of continuous bounded functions. Proof can be found in Stokey et al. (1989), chapter 9 under Theorem,9.6.

A.2. Two period monopoly

From the myopic case we know that in period two an AP will optimally choose to create/redeem $x_2^* = u_2/(2\lambda)$ Thus, in period one a risk-neutral AP is maximizing:

$$V = \max_{x_1} \quad u_1 x_1 - \lambda x_1^2 + \beta E\Big(\frac{u_2^2}{4\lambda}\Big)$$

Assuming $u_2 \sim N\Bigg(\rho\psi(x_1)u_1, \sigma^2\Bigg)$, we can write

$$V = \max_{x_1} \quad u_1 x_1 - \lambda x_1^2 + \beta \frac{(\rho\psi(x_1)u_1)^2 + \sigma^2}{4\lambda}$$

Thus, the FOC (assuming $\psi()$ is a C^0 function) is:

$$x_1 = \frac{u_1}{2\lambda}\Big[1 + \frac{\beta\rho^2 u_1 \psi(x_1)\psi'(x_1)}{2\lambda}\Big]$$

Since the solutions to the FOC are a necessary condition for finding the optimal level of creation/redemption today, x_1, below I characterize some facts regarding them under some general assumptions on the price impact function $\psi(.)$:

1. $\psi(.) \in (0, 1]$

2. $\psi(0) = 1$

3. $sign(\psi'(X)) = -sign(X)$

Proposition 1. *If $u_1 = 0$, then $x1 = 0$*

Proof. Straightforward given the FOC. □

Proposition 2. *If $u_1 > 0$, then $x_1 > 0$. Similarly if $u_1 < 0$, then $x_1 < 0$*

Proof. By contradiction: Suppose $u_1 > 0$. If $x_1 < 0$, then $\frac{u_1}{2\lambda}\Big[1 + \frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big] > 0$ because $\psi\big(x_1\big) > 0$ and $\psi\big(x_1\big) > 0$, thus $\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda} > 0$. As such $\frac{u_1}{2\lambda}\Big[1+\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big]$ cannot be equal to x_1, where $x_1 < 0$, as required by FOC. Hence, if $u_1 > 0$, then $x_1 \geq 0$ Suppose $u_1 < 0$. If $x_1 > 0$, then $\frac{u_1}{2\lambda}\Big[1 + \frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big] < 0$ because $\psi\big(x_1\big) < 0$ and $\psi\big(x_1\big) < 0$, thus $\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda} > 0$. As such $\frac{u_1}{2\lambda}\Big[1 + \frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big]$ cannot be equal to x_1, where $x_1 > 0$, as required by FOC Hence, if $u_1 < 0$, then $x_1 \leq 0$ □

Corollary 3. *Intertemporal arbitraging does not change the sign of the optimal creation/redemption as compared to the myopic choice defined as $\frac{u_1}{2\lambda}$. In particular:*

$$\Big[1 + \frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big] > 0$$

Proof. Follows immediately from Proposition 1 , Proposition 2 and FOC. □

Proposition 4. *Option value of waiting, that arises due to the intertemporal nature of arbitraging, lowers the optimal level of creation/redemption. In particular, for $u_1 \neq 0$ we have:*

$$\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda} < 0$$

Proof. By Proposition 2, if $u_1 > 0$ then $x_1 > 0$. Thus, given the assumptions on price impact function $\psi()$ specified earlier, it has to be that $\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda} < 0$.
Similarly, when $u_1 < 0$ then $x1 < 0$, and, as a result, $\frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda} < 0$ □

Corollary 5. *All solutions to the FOC are between 0 and $\frac{u_1}{2\lambda}$.*

Proof. By Corollary 3 and Proposition 4 we have $0 < \Big[1 + \frac{\beta\rho^2 u_1 \psi\big(x_1\big)\psi'\big(x_1\big)}{2\lambda}\Big] < 1$. Thus, all the solutions to FOC have to be between 0 and $\frac{u_1}{\lambda}$. □

Proposition 6. *If $\psi \in C^1$, then there exist a local maximum at x^* (hence a solution to FOC) between* 0 *and* $\frac{u_1}{2\lambda}$. *Moreover, F attains a global maximum between* 0 *and* $\frac{u_1}{2\lambda}$, *where* $F(x) = u_1 x_1 - \lambda x_1^2 + \beta \frac{(\rho\psi(x_1)u_1)^2+\sigma^2}{4\lambda}$.

Proof. Given that $F'(x) = u_1 x_1 - 2\lambda x_1 + \frac{\beta}{2\lambda}\rho^2 u_1^2 \psi(x_1)\psi\prime(x_1)$, we have $F'(0) = u_1$ and $F'\left(\frac{u_1}{2\lambda}\right) = \frac{\beta}{2\lambda}\rho^2 u_1^2 \psi\left(\frac{u_1}{2\lambda}\right)\psi\prime\left(\frac{u_1}{2\lambda}\right)$.

If $u_1 > 0$, then $F'(0) > 0$ and $F'\left(\frac{u_1}{2\lambda}\right) < 0$. Thus, there exists a local maximum for F between 0 and $\frac{u_1}{2\lambda}$.

Similarly, if $u_1 < 0$, then $F'(0) < 0$ nad $F'\left(\frac{u_1}{2\lambda}\right) > 0$. Thus, there exists a local maximum for F between $\frac{u_1}{2\lambda}$ and 0.

Moreover, by Corollary 3, there are no solutions to FOC outside of the interval between 0 and $\frac{u_1}{2\lambda}$, thus F decreases as x_1 moves away from this interval, and hence F attains a global maximum between 0 and $\frac{u_1}{2\lambda}$. □

A.3. Two period oligopoly

From myopic case we know that in period two each of APs will optimally choose to create/redeem $x^*_{i2} = u_2/(2\lambda^i)$

Thus, in period one each of them will be maximizing:

$$\max_{x_{i1}} u_1 x_{i1} - \lambda^i x_{i1}^2 + \beta E\Big(\frac{u_2^2}{4\lambda}\Big) \tag{B.1}$$

which can be expanded as:

$$\max_{x_{i1}} u_1 x_{i1} - \lambda^i x_{i1}^2 + \beta \frac{(Eu_2)^2 + Var(u_2)}{4\lambda} \tag{B.2}$$

$$\max_{x_{i1}} u_1 x_{i1} - \lambda^i x_{i1}^2 + \beta \frac{(\rho\psi(x_{i1} + X_{-i1})u)^2 + \sigma}{4\lambda} \tag{B.3}$$

FOC: $$u_1 - 2\lambda^i x_{i1} + \tfrac{1}{4\lambda}\Big(2\beta\rho^2 u^2 \psi(x_{i1} + X_{-i1})\psi'(x_{i1} + X_{-i1})\Big) = 0$$

Assuming $\psi(x) = \exp^{-x^2/\phi}$ we have $\psi'(x) = \frac{-2x}{\phi}\psi(x)$

$$u_1 - 2\lambda^i x_{i1} + \frac{1}{4\lambda}\Big(2\beta\rho^2 u_1^2 \frac{-2(x_{i1} + X_{-i1})}{\phi}\psi(x_{i1} + X_{-i1})^2\Big) = 0$$

$$u_1 - 2\lambda^i x_{i1} - \frac{\beta\rho^2 u_1^2 (x_{i1} + X_{-i1})\psi(x_{i1} + X_{-i1})^2}{\lambda^i\phi} = 0$$

Assuming all APs are symetric

$$u_1 - 2\lambda x_{i1} - \frac{-\beta\rho^2 u_1^2 (N x_{i1})\psi(N x_{i1})^2}{\lambda\phi} = 0$$

$$x_{i1} = \frac{u}{2\lambda}\Big(1 - \frac{\beta\rho^2 u_1(Nx_{i1})\psi(Nx_{i1})^2}{\lambda\phi}\Big) \tag{B.4}$$

If $N = 1$, equation B.4 simplifies to the two-period monopoly case. As competition among APs increases, thus as $N \to \infty$, x^*_{i1} approaches myopic case.[26] This is an interesting insight that can help with efficient computation of VFI, also for the infinite horizon case.[27]

[26]Note that by L'Hospital's Rule for any level of x $\lim_{N\to\infty} Nx \exp^{-(Nx)^2/\phi} = 0$. Formal proof utilizes convergence in sequence of function.

[27]Similarly to two-period model, for the infinite horizon case the optimal level of creation/redemption x^*_i will lay somewhere between monopoly infinite horizon case and myopic case.

A.4. Algorithm for Value Function iteration for Bellman oligopoly

We will consider a simplest symmetric model below without freeze constrain.:

1. Set initial guess for policy function $x(u)$ and $V(u)$

2. Perform value function iteration as follows

2a. For each state on the mispricing grid $u^k \in \{u^1, ..., u^{N^u}\}$, using old guesses for V and x compute new guesses $\hat{V}^n$ and $\hat{x}^n$ as follows:

Solve:

$$\hat{x}(u^k) = \arg\max_x CF(u^k, x) + \beta \sum_{j=1}^{N^u} V(u^j) Pr(u^j | u^k, x, X_{-i}(u^k))$$

Update value function:

$$\hat{V}(u^k) = CF(u^k, \hat{x}(u^k)) + \beta \sum_{j=1}^{N^u} V(u^j) Pr(u^j | u^k, \hat{x}(u^k), X_{-i}(u^k))$$

To compute the transition probabilities we need to incorporate competitor's action X_{-i}. We use a previous policy function guess to compute it, in particular

$$X_{-i}(u^k) = (N-1) x(u^k)$$

Transition probabilities between grid points are computed as:

$$Pr\big(u^j | u^k, x, X_{-i}\big) = P\Big(\frac{u^{j-1} + u^j}{2} < u' < \frac{u^j + u^{j+1}}{2} \Big| u^k, x, X_{-i}\Big)$$

transition probabilities for minimum and maximum element on u-grid defined as:

$$Pr\big(u^1 | u^k, x, X_{-i}\big) = P\Big(u' < \frac{u^1 + u^2}{2} \Big| u^k, x, X_{-i}\Big)$$

$$Pr\big(u^{N^u} | u^k, x, X_{-i}\big) = P\Big(\frac{u^{N^k - 1} + u^{N^k}}{2} < u' \Big| u^k, x, X_{-i}\Big)$$

$$\Big(u' | u, x\Big) \sim TN\Bigg(\rho\psi\Big(\sum_i x_i\Big) u, \sigma, \underline{u}, \overline{u}\Bigg)$$

2b. After a complete pass through the grid for u replace old with new guesses $x = \hat{x}$ and $V = \hat{V}$ as in the Gauss-Jacobi scheme, similar to Pakes and McGuire (1994). Alternatively, one could use Gauss-Seidel scheme, in which you replace old with new guesses after visiting each state on u-grid.

3. Iterate with updating until convergence, eg. $\max\left(\left|V^j - V^{j-1}\right|\right)$

A.5. Algorithm for Value Function iteration for Bellman oligopoly with heterogeneous APs and freeze

We assume that there are N authorized participants with heterogeneous parameters (λ^i, γ^i)

1. Set initial guess for policy functions $x_i(u)$ and value functions $V_i(u)$ and $V_i^f(u)$

2. Define random state vector $\omega = (\iota^1, ..., \iota^N)$ where ι^i is a random variable equal 0 if agent i is active today and equal 1 if agent i is frozen today. Assuming noncorrelated freezing constraints, state vector have 2^N combinations of freeze status of APs and the probability of each such state $(\iota^1, ..., \iota^N)$ can be easily computed using γ^i parameters. Let's define a set of all possible states as Ω so that $\omega \in \Omega$

$$Prob\Bigg(\omega = (\iota^1, ..., \iota^N)\Bigg) \ = \prod_{i=1}^{N}(1-\gamma^i)^{1-\iota^i}(\gamma^i)^{\iota^i}$$

2a For each AP i define vector $\omega^{-i} \in \Omega^{-i}$ representing states of competing APs. Thus set Ω^{-i} contains 2^{N-1} possible states and probabilities can be computed as:

$$Prob\Bigg(\omega^{-i}\Bigg) \ = \prod_{j=1, j\neq i}^{N}(1-\gamma^j)^{1-\iota^j}(\gamma^j)^{\iota^j}$$

3. Perform value function iteration as follows

3a. For each state on the mispricing grid $u^k \in \{u^1, ..., u^{N^k}\}$, using old guesses for V_i, V_i^f and x_i compute new guesses $\hat{V}_i$, $\hat{V}_i^f$ and $\hat{x}_i$ as follows:

For each agent $i \in 1, ..., N$ find optimal policy using the V_i, V_i^f and x_i from previous iterations by solving:

$$\hat{x}_i(u^k) = \arg\max_x CF(u^k, x) +$$

$$+\beta \sum_{j=1}^{N^u}\Bigg(\Big[(1-\gamma^i)V_i(u^j) + \gamma^i V_i^f(u^j)\Big]\sum_{m=1}^{2^{N-1}}\Big[Pr(u^j|u^k, x, X_{-i}(\omega_m^{-i}, u^k))Pr(\omega_m^{-i})\Big]\Bigg)$$

To compute the transition probabilities we need to incorporate competitor's action X_{-i}. I denote the join state of all competitors as ω^{-i}

$$X_{-i}(\omega^{-i}, u^k) = \sum_{j=1, j \neq i}^{N} (1 - \iota^j) x_j(u^k)$$

Transition probabilities between grid points are computed as:

$$Pr\Big(u^j | u^k, x, X_{-i}\Big) = P\Big(\frac{u^{j-1} + u^j}{2} < u' < \frac{u^j + u^{j+1}}{2}\Big| u^k, x, X_{-i}\Big)$$

assuming $\Big(u' | u, x, X_{-i}\Big) \sim TN\Big(\rho\psi\Big(x + X_{-i}\Big)u, \sigma, \underline{u}, \overline{u}\Big)$
Transition probabilities for minimum and maximum element on u-grid defined as:

$$Pr\Big(u^1 | u^k, x, X_{-i}\Big) = P\Big(u' < \frac{u^1 + u^2}{2}\Big| u^k, x, X_{-i}\Big)$$

$$Pr\Big(u^{N^u} | u^k, x, X_{-i}\Big) = P\Big(\frac{u^{N^k - 1} + u^{N^k}}{2} < u' \Big| u^k, x, X_{-i}\Big)$$

3b. After a complete pass through the grid for u and for all APs

find updated value functions by computing:

$$\hat{V}_i(u^k) = CF(u^k, \hat{x}(u^k)) +$$

$$+\beta \sum_{j=1}^{N^u} \left(\Big[(1 - \gamma^i) V_i(u^j) + \gamma^i V_i^f(u^j)\Big] \sum_{m=1}^{2^{N-1}} \Big[Pr(u^j | u^k, x, X_{-i}(\omega_m^{-i}, u^k)) Pr(\omega_m^{-i}) \Big] \right)$$

find updated frozen value function V_i^f (eg. solve a system of equations or use fixed point iteration)

$$\hat{V}_i^f(u^k) = \beta \sum_{j=1}^{N^u} \left([(1 - \gamma^i) \hat{V}_i(u^j) + \gamma^i \hat{V}_i^f(u^j)] \sum_{m=1}^{2^{N-1}} \Big[Pr(u^j | u^k, x = 0, X_{-i}(\omega_m^{-i}, u^k)) Pr(\omega_m^{-i}) \Big] \right)$$

3c Check convergence criteria and if not satisfied replace old with new guesses $x_i = \hat{x_i}$ and $V_i = \hat{V_i}$ and $V_i^f = \hat{V_i}^f$ and iterate point 2 and 3 until desired convergence.

BIBLIOGRAPHY

An, Y., Benetton, M., Benetton, M., and Song, Y. (2021). Index providers: Whales behind the scenes of etfs. *SSRN Electronic Journal*.

Antoniewicz, S. R. and Heinrichs, J. (2014). Understanding exchange-traded funds: How etfs work. *ICI Research Perspective*, 20(5).

Antoniewicz, S. R. and Heinrichs, J. (2015). The role and activities of authorized participants of exchange-traded funds. *ICI Investment Company Institute*.

Attari, M. and Mello, A. S. (2006). Financially constrained arbitrage in illiquid markets. *Journal of Economic Dynamics and Control*, 30(12):2793–2822.

Attari, M., Mello, A. S., and Ruckes, M. E. (2005). Arbitraging arbitrageurs. *The Journal of Finance*, 60(5):2471–2511.

Ben-David, I., Franzoni, F., and Moussawi, R. (2017). Exchange-Traded Funds. *Annual Review of Financial Economics*, 9(1):169–189.

Ben-David, I., Franzoni, F., and Moussawi, R. (2018). Do etfs increase volatility? *The Journal of Finance*, 73(6):2471–2535.

Borkovec, M., Domowitz, I., Serbin, V., and Yegerman, H. (2010). Liquidity and price discovery in exchange-traded funds: One of several possible lessons from the flash crash. *The Journal of Index Investing*, 1:24–42.

Brogaard, J., Heath, D., and Huang, D. (2021). The heterogeneous effects of passive investing on asset markets. *SSRN Electronic Journal*.

Brown, D. C., Davies, S. W., and Ringgenberg, M. C. (2020). ETF Arbitrage, Non-Fundamental Demand, and Return Predictability. *Review of Finance*, 25(4):937–972.

Engle, R. and Sarkar, D. (2006). Premiums-discounts and exchange traded funds. *The Journal of Derivatives*.

Evans, R., Moussawi, R., Pagano, M., and Sedunov, J. (2021). Etf short interest and failures-to-deliver: Naked short-selling or operational shorting? *Darden Business School Working Paper*, (2961954).

Fardeau, V. (2021). Strategic trading around anticipated supply/demand shocks. *SSRN Electronic Journal*.

Gromb, D. and Vayanos, D. (2010). Limits of arbitrage. *Annual Review of Financial Economics*, 2:251–275.

Gromb, D. and Vayanos, D. (2018). The dynamics of financially constrained arbitrage. *Journal of Finance*, 73:1713–1750.

Haddad, V., Moreira, A., and Muir, T. (2020). When selling becomes viral: Disruptions in debt markets in the covid-19 crisis and the fed's response. *NBER Working Papers*, (27168).

Harrison, J. M. and Kreps, D. (1978). Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics*, 92(2):323–336.

Hu, H. and Morley, J. (2018). A regulatory framework for exchange-traded funds. *Southern California Law Review*, 91:839–941.

JaneStreetCapital (2018). Comment on the proposed rule 6c-11 under the investment company act of 1940, sec file no. s7-15-18.

Kay, B. (2009). Has the etf arbitrage mechanism failed? why fixed-income etfs traded far from nav, and how the situation has improved. *Morningstar*.

Kondor, P. (2009). Risk in dynamic arbitrage: The price effects of convergence trading. *The Journal of Finance*, 64(2):631–655.

Kozhan, R. and Tham, W. W. (2012). Execution risk in high-frequency arbitrage. *Management Science*, 58(11):2131–2149.

Lettau, M. and Madhavan, A. (2018). Exchange-traded funds 101 for economists. *Journal of Economic Perspectives*, 32:135–154.

Long, J. B. D., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98(4):703–738.

Madhavan, A. (2016). *Exchange-Traded Funds and the New Dynamics of Investing*. Oxford University Press.

Madhavan, A. and Sobczyk, A. (2015). Price dynamics and liquidity of exchange-traded funds. *Journal Of Investment Management*, 14(2):1–17.

Malamud, S. (2016). A dynamic equilibrium model of etfs. *CEPR Discussion Paper No. DP11469*.

Mercadal, I. (2021). Dynamic competition and arbitrage in electricity markets: The role of financial traders. *SSRN Electronic Journal*.

Pakes, A. and McGuire, P. (1994). Computing markov-perfect nash equilibria: Numerical implications of a dynamic differentiated product model. *The RAND Journal of Economics*, 25(4):555–589.

Pan, K. and Zeng, Y. (2019). Etf arbitrage under liquidity mismatch. *Fourth Annual Conference on Financial Market Regulation, Jacobs Levy Equity Management Center for Quantitative Financial Research Pape.*

Petajisto, A. (2017). Inefficiencies in the pricing of exchange-traded funds. *Financial Analysts Journal*, 73(1):24–54.

Shim, J. J. and Todorov, K. (2021). Etfs, illiquid assets, and fire sales. *SSRN Electronic Journal.*

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, 52(1).

Stokey, N. L., Lucas, R. E., and Prescott, E. C. (1989). *Stochastic Dynamic Programming*, pages 239–287. Harvard University Press.

Todorov, K. (2021). The anatomy of bond etf arbitrage. *BIS Quarterly Review, March 2021.*