ETFs, Illiquid Assets, and Fire Sales*

John J. Shim† and Karamfil Todorov‡

October 2023

Abstract

Can ETFs trigger fire sales in illiquid assets? We develop and empirically examine a model where an authorized participant (AP) holds bond inventory and connects the ETF to the underlying bond market. For redemptions, the AP acts as a buffer between the two markets, holding redeemed bonds to preserve the mark-to-market value of her inventory and avoid a fire sale. The AP behaves asymmetrically for creation and transmits ETF purchases to the bond market to boost mark-to-market values. The AP's costs of handling creations/redemptions are paid by liquidity-demanding ETF investors via premiums/discounts. We document new empirical facts motivated by the model, and provide a novel explanation for why ETFs holding more liquid bonds traded at larger discounts than those holding illiquid bonds during the COVID-induced sell-off in March 2020. Our findings show that ETFs have advantages over mutual funds in managing illiquid assets.

Keywords: ETFs, bonds, fire sales, liquidity, COVID

JEL classification: G01, G11, G12, G23

*First version: July 14, 2021. This draft omits some results from the first version (e.g., fractional baskets and their composition), which are included in a separate paper. For more details on the omitted results, please refer to the previous working version currently available on the BIS website or the authors' websites. The views expressed herein are those of the authors and do not necessarily reflect the views of the Bank for International Settlements. We thank Daniel Barth, Jonathan Brogaard, Julio Crego, Caitlin Dannhauser, Prachi Deuskar, Jiekun Huang, Pedro Matos, Jayoung Nam, Stanislava Nikolova, Alex Zhou, and Qifei Zhu who served as discussants. We also thank Zhi Da, Pete Kyle, Marco Pagano, Chris Reilly, Jacob Sagi, Andreas Schrimpf, Sophie Shive, and Russ Wermers for helpful comments. We thank seminar participants at the BIS, Notre Dame, UNC, UTS Sydney, VU Amsterdam, and conference participants at MFA, CFP/Federal Reserve Short-Term Funding Markets Conference, Office of Financial Research of the US Treasury, Future of Financial Information Conference, Notre Dame Institute for Global Investing Conference, Wabash River Finance Conference, Four Corners, FIRS, Professional Asset Management Conference, EEA, the Central Bank Conference on the Microstructure of Financial Markets, the Lapland Investment Fund Summit, and EFA.

†University of Notre Dame Mendoza College of Business, jshim2@nd.edu

‡Bank for International Settlements, karamfil.todorov@bis.org

1 Introduction

Bond exchange-traded funds (ETFs) have grown steadily over the past decade and as of early 2021 manage more than $1.2 trillion compared to less than $220 billion in 2011. This growth has sparked concerns in light of the mismatch between the very liquid ETF market and the less liquid bond market. For example, the famous activist investor Carl Icahn declared that this liquidity mismatch would lead to problems for bond ETFs, stating: "They sell liquidity. There is no liquidity. That's what's going to blow this up." The well known bond investor Bill Gross, dubbed the "Bond King," also articulated a similar concern, stating "the obvious risk – perhaps better labeled the 'liquidity illusion' – is that all investors cannot fit through a narrow exit at the same time."[1] These comments suggest that the mismatch between bond and ETF liquidity may create a problem: if ETF investors sell shares in the liquid ETF market and that selling pressure is transmitted to the less liquid bond market, it could spark a fire sale and hamper the functioning of the bond market. Investors are not the only ones concerned. Regulators and academics have also expressed caution that ETFs could impose systemic risks, especially in periods of market stress.[2]

These concerns about ETF-induced bond fire sales point to an important question: how connected is the ETF market to the underlying bond market? The common narrative is that when an ETF's price deviates from its net asset value (NAV) (i.e., the ETF is trading at a premium or discount), the ETF's authorized participants (APs) engage in arbitrage to correct the discrepancy.[3] This arbitrage takes place through a process called creation/redemption. If the ETF is at a premium to NAV, the AP sells the ETF and delivers a portfolio representative of the ETF's holdings (called the *creation basket*) in exchange for shares of the ETF. If the ETF is at a discount, the AP buys the ETF and redeems it for a representative portfolio of assets *(redemption basket*) which are then sold in the market.

While the creation/redemption process works relatively well in eliminating premiums and discounts for equity ETFs, where both the ETF and the underlying assets are liquid,

[1] Carl Icahn's comments came at the Delivering Alpha conference presented by CNBC and *Institutional Investor* on Wednesday July, 15 2015 (Pramuk, 2015). Bill Gross's comments came in a note to investors in June 2015, as reported by the Wall Street Journal on July 16, 2015 (Zuckerman et al., 2015).

[2] See Bhattacharya and O'Hara (2020) and Pagano et al. (2020) for an overview.

[3] We often use the term "premium" to refer to a positive premium and the term "discount" to refer to a negative premium.

the process is more complicated with bond ETFs. Conceptually, ETF arbitrage could be a double-edged sword when there is a liquidity mismatch between the ETF and underlying assets. On the one hand, if APs tightly connect ETFs and underlying bonds, heavy selling in the ETF is directly transmitted to the less liquid bonds and could potentially trigger a fire sale. On the other hand, if APs are less active in connecting the two markets, selling in the ETF market may have a muted impact on the underlying bonds, allowing discounts to form. In the extreme, if APs withdraw from the market, ETF prices could become unanchored from the value of the underlying bonds.

We develop and test a simple model to understand how liquidity shocks in an ETF that holds illiquid corporate bonds are transmitted to the underlying assets. The model incorporates three important institutional details: (1) corporate bonds are illiquid and selling a large quantity could result in a fire sale, (2) APs hold inventory in corporate bonds, and (3) bond ETFs utilize "fractional" baskets instead of baskets that are proportional to holdings (Shim and Todorov, 2021).[4] The model shows that when ETF investors sell, the AP transmits only part of the selling pressure to the underlying bonds in order to avoid decreasing the mark-to-market value of her inventory.[5] That is, the AP acts as a buffer by absorbing bonds into her inventory instead of selling them in the market, and always prevents fire sales. The AP has an asymmetric response for ETF purchases: she transmits *all* of the buying pressure to the underlying bonds. The AP's costs of holding and trading illiquid bonds are paid by liquidity-demanding ETF investors in the form of premiums and discounts.

We now present more details on the model. When ETF investors sell for liquidity reasons, the AP buys the ETF from them at a discount and redeems the ETF for a basket of bonds. The AP can do a combination of two things with the redeemed bonds: sell them, which pushes down bond prices, or add them to her inventory and avoid reducing bond prices. In equilibrium, the AP holds at least some fraction of the redeemed bonds to protect the mark-

[4]Bond ETF baskets typically only include about 10%-40% of the bonds that are held by the ETF. Shim and Todorov (2021) show that the use of such "fractional baskets" is one way in which bond ETFs manage illiquid assets.

[5]When we use the term AP, we refer to any bond dealer that trades in both the ETF and bond markets. In practice, these dealers may not be officially listed as an AP, but they commonly utilize the creation/redemption mechanism through registered APs.

to-market value of her inventory, and always avoids selling an amount that would trigger a fire sale. That is, the AP acts as a buffer between the ETF market and the bond market and avoids depressing bond prices because she has "skin in the game."[6]

When ETF investors buy the ETF and trigger a creation, the AP responds differently compared to redemption. Instead of selling only a *fraction* of the bonds in the market as in the case of a redemption, the AP always buys *all* of the creation basket bonds in the market. The AP does so for two reasons: (1) either she does not have the basket bonds in her inventory, or (2) she holds the bonds but chooses to buy them in the market, in an effort to boost the mark-to-market value of her inventory. The behavior of the AP in creations vs. redemptions highlights an asymmetry: ETF purchases are always transmitted in full to the underlying bonds, whereas ETF sales are not.

The model also provides an interpretation of ETF premiums and discounts as a transaction cost for liquidity-demanding ETF investors. The conventional wisdom around ETFs is that they are very liquid, a conclusion supported by their narrow bid-ask spreads. However, this view misses an important point: when ETF investors trade, they tend to buy at a premium or sell at discount (also articulated in Koont et al. (2022)).[7] In fact, the model illustrates that premiums and discounts materialize exactly because of liquidity-demanding ETF investors and reflect the AP's cost of handling creation/redemption bonds. Since corporate bonds are costly to trade and hold, the premiums and discounts can be significant.

The model also illustrates that ETFs have an important advantage in managing illiquid assets over mutual funds. In the case of a mutual fund, redeeming investors trigger forced selling of bonds to meet redemptions, even if that leads to a fire sale. This forced selling results in lower NAV for non-redeeming mutual fund investors, and imposes an externality on the bond market as a whole. As a precaution, mutual funds may hold cash or more liquid assets to avoid forced selling of illiquid assets, but doing so simply transforms liquidation

[6]This result relies on the assumption that APs can preserve mark-to-market values by not selling bonds, which captures real-world sentiment by APs. The chief executive of one ETF market-making firm was aptly quoted as saying, "If the bonds don't trade then you don't have to write them down" (Johnson, 2023).

[7]The source of premiums and discounts in Koont et al. (2022) comes from the AP's cost of liquidity transformation, and ETFs include cash in creation/redemption baskets to reduce the cost. In our model, discounts stem from the AP's costs of avoiding the liquidation of illiquid assets to prevent mark-to-market losses and fire sales; premiums arise from trading costs and mark-to-market gains.

costs into tracking error.[8] In contrast, large ETF redemptions do not trigger a fire sale because the AP acts as a buffer. Since the AP internalizes the cost of lower bond prices, her incentives are aligned with bond investors and non-redeeming ETF investors. That is, a mutual fund manager acts as a liquidation agent on behalf of *redeeming* investors, which could impose externalities on remaining investors and the bond market. In contrast, the AP acts as a liquidation agent on behalf of *bond* investors (including remaining ETF investors), and redeeming investors pay the liquidation cost in the form of discounts.

The AP's role as a buffer between the ETF and bond markets is possible because of an important design feature of ETFs: in-kind transfers. When the AP buys the ETF at a discount and redeems it for bonds, she effectively buys *bonds* at a discount and does so without engaging the secondary bond market. This means that bond prices are not updated, and the acquisition of the bonds does not trigger a need to update bond mark-to-market values. The decoupling of bond trades and mark-to-market values is what allows the AP to effectively act as a buffer. If a mutual fund were to raise cash by selling bonds, it would necessarily link bond transaction prices and mark-to-market values: a dealer either buys bonds at low prices and marks down existing inventories, or preserves mark-to-market values but does not buy the bonds at a discount.

We also provide empirical support for the model. We introduce a new methodology to infer creation-redemption baskets based on daily ETF holdings, which allows us to identify creation and redemption bonds with high accuracy. The first empirical fact that lends support to our model is that the abnormal dollar trading volume in creation bonds lines up almost one-for-one with the dollar creation amount. This finding is consistent with the model's prediction that APs purchase all creation bonds in the bond market. We also find that the abnormal trading volume in redemption bonds is consistent with the model's prediction that the AP trades only a fraction of the redeemed bonds in the market. We estimate that APs only trade roughly 15%-20% of the redeemed bonds and hold the remainder.

The second empirical fact we document is that discounts are increasing in APs' inventories. This fact is consistent with the model's prediction that greater inventory leads to

[8]Mutual funds can also use redemption gates or rely on a mechanism called swing pricing, which allows them to adjust the price at which investors redeem by a swing factor. See Section 3.6 for details on how ETF discounts can be used to calibrate the swing factor in real time.

a larger discount since the AP holds a greater fraction of redeemed bonds (i.e., she has more skin in the game), and the costs of holding more bonds are passed to redeeming ETF investors.

Third, we show that the pattern of returns after creations and redemptions is consistent with the model's interpretation of premiums and discounts as trading costs to ETF investors. We examine cumulative ETF and NAV returns following a creation/redemption day with a premium/discount. We find that premiums (positive or negative) negatively forecast future ETF returns, but are unrelated to future NAV returns. These findings are consistent with our model, which predicts that ETF prices should revert following the liquidity shock from ETF investors, but NAV should have a muted response.

Finally, our model provides an explanation for some of the puzzling facts observed during the COVID-induced market stress of March 2020. This episode provides a particularly interesting test for our model since bond ETFs saw significant redemptions that had the potential to trigger a fire sale. A commonly-held view is that ETFs holding the most illiquid bonds should have the deepest discounts in times of stress since investors may be more likely to avoid trading illiquid bonds, leaving prices of those bonds stale (Madhavan and Sobczyk, 2016). However, our model predicts that the ETFs holding bonds that are most concentrated in APs' inventories should see the deepest discounts as APs avoid selling those bonds to protect mark-to-market values. Since APs hold the largest inventory in investment-grade bonds (especially short maturity) and little to no inventory in high-yield bonds, our model predicts, perhaps surprisingly, that investment-grade ETFs should see larger discounts than high-yield ETFs.

The data on discounts during March 2020 support our model's prediction: the ETFs that traded at the deepest discounts were those holding investment-grade bonds (especially short-maturity). For example, the average short-maturity investment-grade ETF traded at a discount above 6%. This is in contrast to ETFs holding riskier and less liquid high-yield bonds, which had average discounts *2 to 6 times smaller* than more liquid investment-grade ETFs. This fact is puzzling based on the commonly held view, which argues that illiquidity and stale prices cause discounts, but is consistent with our novel inventory-based

explanation.[9]

The remainder of this paper is organized as follows. We discuss related literature below. Section 2 describes important ETF institutional details, Section 3 presents the model, Section 4 describes the data and provides empirical support for the model's predictions, and Section 5 concludes.

Related Literature Our paper is related to three distinct strands of literature: ETFs, bond market illiquidity, and fire sales.

First, this paper is related to the growing literature on ETFs. Ben-David et al. (2018) show that ETF arbitrage transmits noise from the ETF market to the underlying securities and increases volatility. Malamud (2015) demonstrates that ETFs can create a transmission mechanism for non-fundamental shocks to underlying securities. Saglam et al. (2019) show that ETFs increase stock liquidity, and Da and Shive (2018) find that they increase return comovement. Brogaard et al. (2019) show that the impact of ETFs on liquidity depends on the fund's index replication strategy. Shim (2020) argues that ETF arbitrage mistranslates systematic information from ETFs to constituent securities. Todorov (2019) shows that ETFs in VIX and commodity markets put pressure on prices and amplify price changes. Sushko and Turner (2018) document the increase in the share of ETFs in several markets and study the impact for liquidity and volatility.

Several papers analyze bond ETFs in particular. Dannhauser (2017) finds that ETF ownership increases prices but reduces bond liquidity. Holden and Nam (2019) find that ETFs improve liquidity in less accessible corporate bonds. Pan and Zeng (2021) show that APs' dual role as major bond dealers and ETF arbitrageurs means that balance sheet constraints can affect arbitrage. Reilly (2021) argues that APs tend to include underperforming bonds in creation baskets. Koont et al. (2022) describe the tradeoff that ETFs face in managing illiquid assets, where liquid assets or cash can act as a substitute for more illiquid assets. In addition,

[9]Haddad et al. (2021) document that investment-grade corporate bond spreads traded at a discount to credit default swaps. The authors suggest that this effect was driven by a desire by mutual funds and other liquidity-demanding bond investors to liquidate large positions. Ma et al. (2022) document that bond mutual funds sold Treasuries and more liquid corporate bonds to meet redemptions. While both of these papers paint a picture of depressed corporate bond prices, they do not explain why investment-grade *ETFs* traded at deeper discounts. In fact, selling pressure in bonds would suggest that investment-grade ETFs are *less* likely to trade at a discount as bond returns might outpace ETF returns.

Koont et al. (2022) also show that ETF premiums and discounts can be interpreted as costs. However, their channel is distinct and complimentary to our paper. Their model shows that premiums and discounts can be made smaller by including cash in baskets, whereas our model shows that premiums and discounts come from the AP's costs of preventing fire sales and managing the mark-to-market value of inventory. Shim and Todorov (2021) document that bond ETFs use fractional baskets, which results in persistent premiums and discounts. Haddad et al. (2021), D'Amico et al. (2020), Aramonte and Avalos (2020), and Boyarchenko et al. (2020) study bond ETFs during the COVID crisis and the subsequent Fed intervention. Todorov (2021) describes the specifics of the bond ETF arbitrage mechanism.

Our paper is also related to a growing literature studying whether ETFs are a source of systemic risk. Ramaswamy (2011) argues that ETFs may introduce systemic risk due to the greater interconnectedness of products. Bhattacharya and O'Hara (2018) demonstrate that ETFs holding illiquid assets may distort prices of underlying assets due to challenges in discerning which shocks are pertinent to which assets. Pagano et al. (2020) and Bhattacharya and O'Hara (2020) provide an overview of the ETF literature with a specific focus on periods of market disruption and examine whether ETFs increase systemic risk.

Our research is also related to studies of bond market liquidity. Bessembinder et al. (2020) provide a summary of the main differences between equity and fixed income markets. Bessembinder et al. (2018) show that the median trade size in bond markets is more than $1 million, which is much larger than that in equity markets. Bessembinder et al. (2009) find that round-lot trades account for approximately 90% of corporate bond dollar trading volume. Goldstein and Hotchkiss (2020) show that dealers endogenously adjust their behavior to mitigate inventory risk from trading in illiquid and higher-risk bonds. Bao et al. (2011) show that the illiquidity of corporate bonds impacts their prices. O'Hara and Zhou (2021) illustrate that primary dealers played a mostly positive role for liquidity provision during the COVID-19 stress.

The research presented here also contributes to the extensive literature on fire sales. This literature stems from the classic paper of Shleifer and Vishny (1992), which shows that forced sales can lead to depressed liquidation values and can have significant implications for firms' cost of capital. Coval and Stafford (2007) illustrate that mutual fund outflows can lead to fire

sales and financial distress in equity markets. Ellul et al. (2011) study fire sales of corporate bonds induced by regulatory constraints imposed on insurance companies. Goldstein et al. (2017) show that the illiquidity of corporate bonds may generate a first mover advantage in corporate bond mutual funds and amplify price moves, whereas Choi et al. (2020) find little evidence that corporate bond mutual fund redemptions cause fire sales.

2 ETF Institutional Details

There are two markets for ETFs: a primary market, where new ETF shares are created and destroyed, and a secondary market, where investors trade existing ETF shares. In the primary market, Authorized Participants (APs) create new or redeem existing ETF shares in exchange for a portfolio of assets (bonds or stocks); this portfolio is referred to as the creation/redemption basket. Importantly, as shown in Shim and Todorov (2021), this basket may not contain all of the securities held in the ETF's portfolio and can utilize different portfolio weights compared to holdings. The secondary market is where market participants trade ETF shares, just like they trade stocks. APs are usually large broker-dealers or liquidity providers in the underlying assets and may operate both in the primary and secondary market. In addition, many market participants that engage in AP activity may not be listed as APs, but may rent an AP's "pipes" to create and redeem ETF shares.

Conceptually, the process of creation/redemption works because APs are incentivized to eliminate ETF premiums/discounts through arbitrage profits. The textbook version of this arbitrage is as follows. During the course of a typical trading day, price pressure may cause the ETF price to deviate from the NAV of its underlying assets. APs can then buy the asset with the lower price and sell the asset with the higher price until the two converge. APs do so because they are able to use the creation/redemption mechanism to convert the portfolio of ETF holdings into ETF shares and vice-versa. This conversion nets out the AP's position and locks in arbitrage profits from buying low and selling high.

A common assumption in ETF arbitrage is that ETF holdings and creation/redemption baskets are identical. This ensures that when APs eliminate discrepancies between the ETF price and the NAV of the *basket*, they also eliminate discrepancies between the ETF price

and the NAV of the *holdings*. The assumption that baskets are equal to holdings is generally true for equity ETFs (Todorov, 2021) but not for bond ETFs, which use fractional baskets (Shim and Todorov, 2021). The fact that bond ETF baskets are a fraction of holdings is likely because bonds are less liquid than equities, which makes trading the entire portfolio of bond holdings costly. As a result, it is relatively common for bond ETF sponsors to create so-called custom baskets. These baskets are typically different from holdings as well as from the pre-announced basket that ETF sponsors publicly declare but infrequently adjust. Custom baskets are an important element of ETFs holding relatively illiquid assets. They allow for more flexibility in the creation/redemption process by taking trading frictions and inventory constraints into account, but lead to more persistent premiums and discounts (Shim and Todorov, 2021).

It is common that bond ETFs trade at a premium or discount as we show in Section 4.1.3, which illustrates that the arbitrage mechanism does not perfectly align bond ETF prices with NAV. We next turn to the model and show that an AP's inventories and costs are an important factor that could explain the price discrepancies between an ETF and its underlying bonds.

3 Model

We build a simple model to understand how trading in the ETF market transmits to the underlying bonds. The central agent in our model is an AP who holds inventory in bonds as a dealer in the bond market, acts as a liquidity provider in the ETF market, and engages in arbitrage between the two markets. The model helps understand how the AP's tendency to hold inventory in the underlying bonds affects her choice of what price to provide ETF liquidity, the determinants of premiums and discounts, and the degree of transmission of ETF shocks to the underlying bonds.

We analyze redemptions (ETF sales) and creations (ETF purchases) separately in the model, and show that the AP behaves in an asymmetric fashion for creations vs. redemptions. We start by describing the model setup and equilibrium for redemptions, then the setup and equilibrium for creations. We then discuss some model extensions, compare the ETF to a

stylized mutual fund, and present four testable hypotheses based on our model.

3.1 Redemption Model Setup

There are three securities in our model: two bonds, denoted A and B, and an ETF that holds an equally-weighted portfolio of the two bonds. Initially, the price of the bonds and the ETF are all equal to P_0. The NAV of the ETF is given by the weighted average of the holdings, i.e., initially it is equal to the ETF price P_e ($NAV = P_e = P_0$). Motivated by the use of fractional baskets by bond ETFs, we initially assume that the ETF basket contains only one of the two bonds (we present a version of the model where the basket contains both bonds ("full baskets") in Appendix B.1).

There are two agents in the model. The first is an uninformed investor who exogenously needs to sell the ETF in the secondary market. The second is a representative risk-neutral AP who provides liquidity in the ETF market by trading with the investor (we extend the model to include multiple APs in Section 3.5). The AP is representative in the sense that she acts as if there is perfect competition among many APs to provide liquidity to the uninformed investor in the secondary ETF market. Specifically, the AP buys from the investor at the maximum ETF price that yields zero profits. In addition to the secondary ETF market, there are two other markets: a primary ETF market where the AP redeems ETF shares for the redemption basket, and a bond market where individual bonds are traded.

There is a single trading period, in which the investor submits a market order in the secondary ETF market to sell $0 < q < 0.5$ units of the ETF, where q is the amount sold as a fraction of the total AUM of the ETF.[10] The AP sets the ETF price at which she is willing to buy the q units of the ETF from the investor. We also assume that the AP then immediately redeems all of the newly purchased ETF units for the redemption basket and adds these bonds to her inventory position.[11]

The AP initially holds an equal amount of inventory z in each bond A and B, i.e.,

[10] In other words, the initial AUM of the ETF is the numeraire.

[11] In reality, APs may hold inventory in both the ETF and bonds instead of only bonds. For example, instead of redeeming all ETF shares, the AP may redeem only a fraction of the shares in exchange for the redemption basket. To simplify the analysis, we assume that the AP always redeems bonds instead of holding units of the ETF without redeeming. This allows us to analyze inventory holdings of only bonds instead of inventory in both bonds and ETFs.

$z_A = z_B = z \geq 0$. We assume the initial inventory is ex-ante optimal given the inventory needs of the AP as a bond dealer (Pan and Zeng, 2021). The AP buys $q > 0$ units of the ETF in the secondary market at price P_e, determined below, and receives bonds in the redemption basket worth NAV_{basket}. The AP then chooses to keep a fraction γ of the redeemed bonds in her inventory and sells the remaining fraction $1 - \gamma$ in the bond market.

The AP incurs three types of costs associated with holding and trading the redeemed bonds. First, she faces *price impact costs* when trading bonds. When she sells a bond, the transaction price of that bond is given by $P_i - C((1-\gamma) \cdot q)$, where i denotes the bond (A or B) and the price impact function $C(x)$ is linear in the quantity of bonds traded. In addition, the bond's transaction price instantaneously triggers an update to the mark-to-market value for those holding the bond. We capture the illiquidity of bonds in the spirit of Shleifer and Vishny (1992) by assuming that the bond market can only absorb a certain quantity of bonds up to a threshold τ. Trading more than τ leads to a fire sale in the sense that the impact on bond prices is significantly greater. Specifically, the price impact cost is given as

$$C(x) = \begin{cases} cx, & x \leq \tau \\ fx, & x > \tau, \end{cases}$$

where $f > c > 0$, and f is the marginal cost in the case of a fire sale. Thus, selling more than τ leads to a discontinuous jump in price impact. This assumption captures the idea that if the AP trades bonds in "typical" amounts, she is able to find "specialist" counterparties who charge lower transaction costs, i.e., are willing to pay higher prices. However, if the AP needs to trade bonds in large amounts (i.e., above τ), she needs to trade with "non-specialists" who offer lower prices for the bond (and thus the AP realizes greater price impact costs).[12]

The notion of "specialists" builds on Shleifer and Vishny (1992), where, in times of fire sales and when specialists are constrained, distressed firms need to sell their assets to outsiders who have lower valuations.[13] The discontinuity in the cost function at τ comes

[12]This idea is consistent with BlackRock's own description of the types of APs that participate in the primary market. In a webinar to discuss ETF primary market activity held on May 10, 2021, BlackRock stated that dealers specializing in particular types of bonds were more likely to conduct primary market activity than non-specialized dealers, especially in recent years (BlackRock, 2021).

[13]In our model, we assume that there is only one specialist buyer: the AP cannot split the large amount

from the assumption that the price impact cost of all units (including inframarginal units below τ) is f. There are two reasons for this. First, this specification captures the possibility that specialist buyers are constrained in times of fire sales and the AP has to sell the whole amount to non-specialists. Second, the specification is a reduced form way of capturing a jump at τ that comes from greater fixed costs associated with units sold to non-specialists.[14] Fire sales in our model can also be viewed as representing the case where there are significant negative externalities to bond investors and the bond market in general.[15] We note that our main findings related to premiums/discounts and inventories do not rely on the discontinuity in the price impact function.

The second cost the AP faces is a *holding cost* of additional inventory. We assume the cost of increasing holdings by x is $\frac{\lambda}{2}x^2$ with $\lambda > 0$. Since we assume that the AP's initial inventory level z is optimal, these costs arise due to *deviations* from this level. This specification of holding costs is similar to the one in Pan and Zeng (2021) and Koont et al. (2022), and captures the idea of corporate bond inventory costs (Goldstein and Hotchkiss, 2020; Bessembinder et al., 2018) and balance sheet capacity costs (Andersen et al., 2019). We make an additional restriction that λ is bounded above by $\bar{\lambda}$ (the formal expression is provided in Appendix Section A.1). If λ is above this bound, holding costs are so large that the AP always prefers to sell, even at fire sale prices, rather than to hold redeemed bonds. Since fire sale costs are usually thought to be relatively large, this restriction allows us to focus on cases where fire sales impose a significant cost, i.e., avoiding them could provide a significant positive externality. The case where λ is above $\bar{\lambda}$ may represent a scenario where APs are so constrained that they are unwilling to do anything except sell the bonds and act as a pass-through between selling in the ETF market and selling in the bond market.

The third type of cost that the AP incurs is a *mark-to-market cost*, which arises from adjustments to the valuation of the AP's bond inventory. If a bond is traded in the market, its market price is updated, which triggers an update in valuation. This is the sense in which

to several buckets of size τ and sell them to different specialist buyers.

[14] Another way to interpret this function is that the marginal cost of additional units sold above τ is decreasing. An alternative specification is to make the marginal cost of all bonds sold after τ constant, in which case the cost would be $f+c$. We adopt the simpler version of the cost since the alternative specification adds significant complexity without changing the main intuition.

[15] This is why convex price impact, for example, may not capture the discontinuity between normal trading activity and large sales which trigger broader contagion.

the AP internalizes the cost of lower bond prices. Since the AP holds some initial inventory in bonds, lower bond prices decrease the mark-to-market value of the AP's bond holdings. This cost represents a decline in the AP's borrowing capacity or equivalently, an increase in leverage and a decrease in capital ratios. We model this cost as the total change in value of the initial bond inventory plus the newly acquired bonds that are held after redemption, or $(z + \gamma q)\Delta P_i$.

3.2 Redemption Equilibrium

In the spirit of fractional baskets, we assume that the redemption basket only contains bond A. In the case of an ETF redemption of q shares, the AP earns an arbitrage profit of $NAV_{basket} - P_e$, which is equal to $P_0 - C\left((1-\gamma)\,q\right) - P_e$ per unit. That is, the arbitrage profit is the difference between the price after selling bonds in the bond market (initial NAV minus price impact) and the purchase price of the redeemed bonds (i.e., the ETF price).[16] The AP's total payoff is

$$\underbrace{(NAV_{basket} - P_e) \cdot q}_{\text{arbitrage profit}} - \underbrace{\frac{\lambda}{2} \cdot (\gamma q)^2}_{\text{holding cost}} + \underbrace{(z + \gamma q) \cdot \Delta P_A}_{\text{mark-to-market cost}} \tag{3.1}$$

where $\Delta P_A = -C\left((1-\gamma)\,q\right)$.[17]

The AP's equilibrium choice of γ is different depending on whether redemptions are large enough to trigger a fire sale, or not. We first analyze the case when redemptions are "typical" and do not trigger a fire sale. In that case, we initially assume that price impact costs are c regardless of the traded quantity. We denote γ_c as the fraction of bonds held (the fraction

[16] An alternative specification for arbitrage profits allows the number of ETF shares in creation/redemption to be determined by the NAV of holdings (as is often the case in reality), instead of the NAV of the basket (as specified above). This alternative specification is $(NAV_{basket} - kP_e)q$, where $k = q\frac{NAV_{basket}}{NAV} = q\frac{P_0 - c(1-\gamma)q}{P_0 - 0.5c(1-\gamma)q} < 1$ is the number of ETF shares that the AP delivers to get one unit of NAV_{basket}. In other words, instead of redeeming one ETF share to get one unit of NAV_{basket}, the AP redeems $k < 1$ shares valued at NAV, which increases her arbitrage profit. This specification of arbitrage profits significantly complicates the derivation of the optimal γ (as k depends on γ and hence the optimal γ depends on P_e, which makes the solution more algebraically complex) without changing the main intuition. Since empirically $k \approx 1$, we present the results with the much simpler version of the model, where the AP delivers one ETF share per unit of NAV_{basket}.

[17] Given that creation/redemption fees are typically very small, we assume they are zero for simplicity. The logic is unchanged with constant non-zero fees.

sold is then $1-\gamma_c$) when price impact costs are fixed at c. We then analyze the second case, when redemptions are large enough to trigger a fire sale. In that case, if the AP were to sell the fraction of bonds acting as if the price impact cost was c, she would sell more than τ and incur a price impact cost of f rather than c. That is, employing the "typical" redemption selling strategy would lead to a fire sale. To avoid the fire-sale scenario, the AP revises the choice of γ to take into account the significant increase in price impact costs, adjusting it from γ_c to the fire-sale fraction denoted γ_τ. That is, the AP chooses to hold more redeemed bonds ($\gamma_\tau > \gamma_c$) to endogenously avoid a fire sale. We now describe each of the two cases.

Case 1: No Fire Sales

In this case, the AP incurs price impact costs of $c < f$ when selling bonds (for now, even if the selling quantity exceeds τ, we assume that price impact costs are still c). Since the representative AP is disciplined by perfect competition, we set her payoff from equation 3.1 to zero. Rearranging equation 3.1 yields an expression for the ETF price P_e as a function of the AP's choice of γ:

$$P_e(\gamma) = P_0 - c(q+z) + cz\gamma - \left(\frac{\lambda}{2} - c\right) q\gamma^2. \tag{3.2}$$

In keeping with perfect competition, the AP selects γ such that she maximizes the ETF price that ETF investors receive.[18]

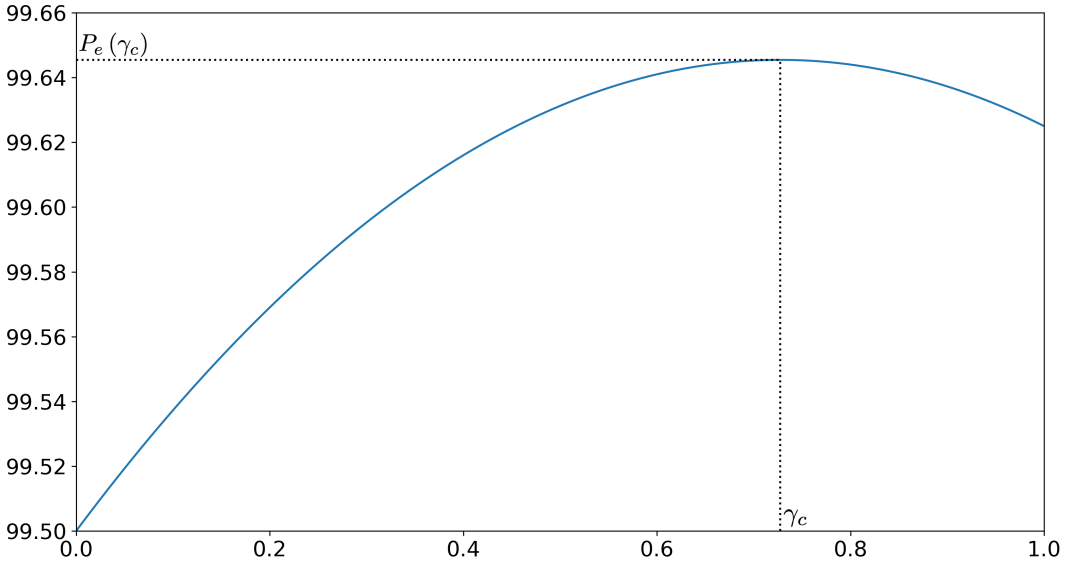
There are two key conditions for the solution γ_c to be interior, i.e., γ_c between 0 and 1. First, holding costs $\left(\frac{\lambda}{2}\right)$ must be greater than the non-fire sale price impact cost (c). Second, the AP's inventory z must be below the threshold $q\frac{(\lambda-2c)}{c}$. If $\frac{\lambda}{2} \leq c$ (holding costs are too low) or $z \geq \frac{\lambda-2c}{c}q$ (inventory is too large), the AP holds all bonds (i.e., $\gamma_c = 1$). Otherwise, γ_c is interior and is given by

$$\gamma_c = \frac{cz}{(\lambda - 2c)\, q}, \tag{3.3}$$

and is increasing in the ratio of the initial inventory z to the redeemed quantity q. This

[18]The logic is consistent with market clearing in a competitive market in which an investor sells q in the secondary ETF market and APs compete to purchase q, which yields the highest price for the investor and no profits for APs. It is straightforward to extend the model with non-zero AP profits by adding an additional term that is constant or scales with q but not γ in equation 3.2.

Figure 3.1: Typical Redemption



Notes. This figure plots the ETF price $P_e(\gamma)$ as a function of γ (the share of redeemed bonds that the AP puts on her balance sheet) and shows the equilibrium holding fraction when price impact costs are fixed at c: γ_c. For this example, we use $q = 0.05$, $z = 0.2$, $c = 2$, $f = 4$, $\lambda = 15$, $\tau = 0.1$, and $P_0 = 100$. See Section 3.2 for more details on the shape of the function and the conditions for an interior γ_c.

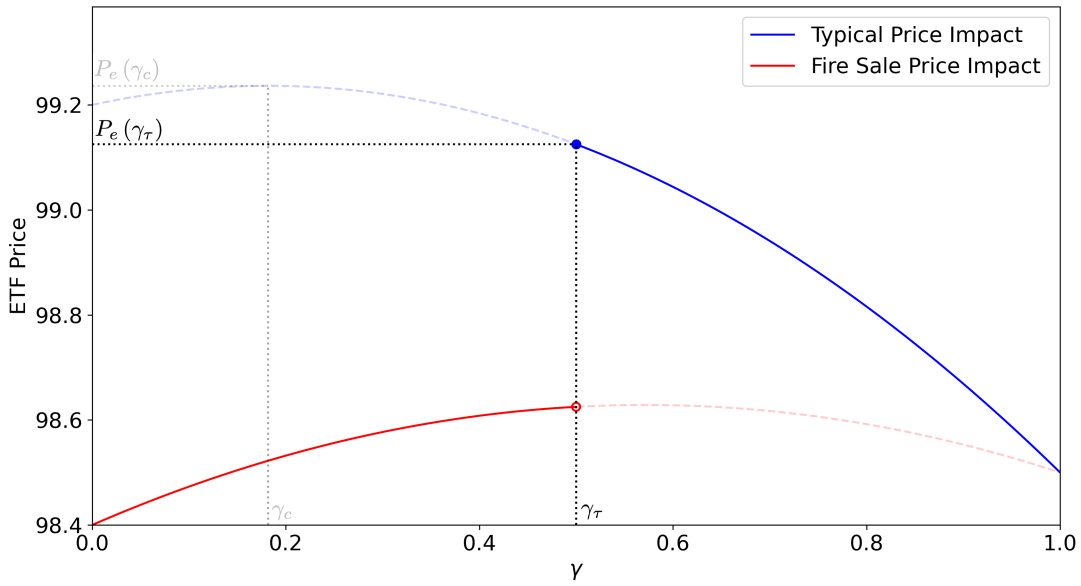
captures the tradeoff that the AP faces: reducing mark-to-market losses (which are increasing in z) vs. reducing excess holding costs (which are increasing in q). The larger the inventory relative to the redeemed amount, the larger the fraction of redeemed bonds that the AP decides to hold to avoid the cost from lower bond prices. Figure 3.1 provides an example of an interior γ_c and the resulting ETF price in the case of a typical (non-fire sale) redemption.

The characterization of the AP's solution in this non-fire sale case depends on the redemption quantity q being sufficiently small such that the AP's optimal selling quantity $(1 - \gamma_c)\, q$ is smaller than the fire-sale threshold τ. If this condition is satisfied, the AP can select γ_c since the selling quantity does not exceed the fire-sale threshold. If this condition is not satisfied, there is a fire sale and the price impact costs jump from c to f. As a result, the ETF price given by γ_c is not achievable.

Case 2: Fire Sales, $(1 - \gamma_c)q > \tau$

We now consider the fire sale case when $(1 - \gamma_c)q > \tau$. In the fire sale region, the AP considers adjusting her holding fraction from γ_c to an alternative fraction which we denote

Figure 3.2: Fire Sale Redemption



Notes. This figure plots the ETF price $P_e(\gamma)$ as a function of γ (the share of redeemed bonds that the AP keeps on her balance sheet) in the case when $(1-\gamma_c)q > \tau$. The solid lines represents the ETF price, which is discontinuous when the selling quantity $(1-\gamma)\,q$ is greater than the fire sale threshold τ. The equilibrium holding fraction in the fire sale scenario γ_τ is the minimum amount held such that the price impact cost is still $c < f$. For this example, we use $q = 0.2$, $z = 0.2$, $c = 2$, $f = 4$, $\lambda = 15$, $\tau = 0.1$, and $P_0 = 100$. See Section 3.2 for more details.

γ_τ. The alternative fraction γ_τ must take into account the increase in price impact costs from c to f when selling more than τ. Note that the redeemed quantity q can be greater than τ but a fire sale can still be avoided if $(1-\gamma_c)\,q < \tau$. This is possible since the AP sells only a fraction of the redeemed bonds.

Since for each possible value of $\gamma \in [0, 1)$, the ETF price is greater with price impact costs c than with costs f, the AP considers γ_τ, which is the smallest holding fraction that satisfies $(1-\gamma_\tau)\,q \leq \tau$, or $\gamma_\tau = 1 - \frac{\tau}{q}$. That is, the AP maximizes the ETF price by selling up to the threshold τ in order to avoid the discontinuous jump in price impact costs, and the resulting drop in the ETF price. The intuition behind this choice is simple. Since the AP holds initial inventory in the redeemed bonds, she internalizes the cost of fire sales. In addition, even in the absence of inventory, competition pushes the AP to avoid a fire sale because she can economize on costs and she is able to quote a higher ETF price.

Figure 3.2 provides an example of the case where the AP adjusts her holding fraction to avoid a fire sale. The graph shows that γ_c would yield the highest ETF price but is not

achievable because selling $(1-\gamma_c)q$ triggers a jump in price impact costs from c to f and a drop in the ETF price. As a result, the AP chooses γ_τ, the smallest γ possible to avoid a fire sale and retain price impact costs of c. Combining the two cases above allows us to describe the equilibrium.

Proposition 1. *(Redemption with Fractional Baskets) When $\lambda \leq 2c$ (bond holding costs are low) or $z \geq \frac{\lambda-2c}{c}q$ (inventory is high), the equilibrium fraction of the redeemed bonds that are held by the AP in inventory, γ^*_{RD}, is 1. Otherwise, γ^*_{RD} is given by*

$$\gamma^*_{RD} = \begin{cases} \gamma_c, & \text{if } (1-\gamma_c)q \leq \tau \text{ ("typical" redemption)} \\ \gamma_\tau, & \text{if } (1-\gamma_c)q > \tau \text{ ("fire sale" redemption),} \end{cases}$$

where

$$\gamma_c = \frac{cz}{(\lambda - 2c)\, q}$$
$$\gamma_\tau = 1 - \frac{\tau}{q}.$$

The NAV of the ETF and the ETF price are

$$NAV^*_{RD} = P_0 - \frac{0.5-q}{1-q}(1-\gamma^*_{RD})cq$$
$$P^*_{e,RD} = P_0 - c(q+z) + \gamma^*_{RD}cz - (\gamma^*_{RD})^2\left(\frac{\lambda}{2} - c\right)q,$$

and the ETF discount is

$$-\pi^*_{RD} \equiv NAV^*_{RD} - P^*_{e,RD} = (\gamma^*_{RD})^2\left(\frac{\lambda}{2} - c\right)q + (1-\gamma^*_{RD})\,cz + \frac{0.5 + \gamma^*_{RD}\,(0.5-q)}{1-q}cq. \quad (3.4)$$

3.2.1 Equilibrium Discussion

We discuss three important observations that follow from Proposition 1. First, the AP's optimal holding fraction γ^*_{RD} always avoids selling quantities that lead to a fire sale. The intuition for this result is straightforward – the AP internalizes the cost of a fire sale because

she holds inventory in the redeemed bonds, and thus triggering a fire sale would result in costly mark-to-market losses. The fact that the AP acts as a shock absorber suggests that ETFs can be an important mechanism to insulate illiquid bonds from fire sales. Perhaps the most significant economic consequence is that firms can access debt markets in periods of stress at lower cost, since bond markets can continue to function relatively normally compared to the scenario where large ETF redemptions propagate to the underlying bonds.

Second, we assume that competition disciplines the AP and results in her setting the highest ETF price possible, but this does not necessarily minimize the ETF discount. Equation 3.4 shows that the discount in a typical, non-fire sale redemption is minimized when γ is equal to $\frac{c}{\lambda-2c}\frac{z}{q} - \frac{c(0.5-q)}{(1-q)(\lambda-2c)} < \gamma_c$. That is, the γ that minimizes the discount is smaller than the one that maximizes the ETF price. This is an important takeaway from our model, which shows that when the AP has non-zero inventory, the optimal choice of γ does not necessarily minimize the discount. This finding is contrary to conventional wisdom. In the classical arbitrage case (as in equity ETFs), the AP completely eliminates ETF discounts (when $P_e < NAV$). In our model, this is only true in the case with full baskets and where $z = 0$ (see Proposition 4 in Appendix B.1). However, the model highlights that price impact costs affect the mark-to-market value of the AP's inventory and prevent the AP from eliminating the discount completely. That is, with inventory, the AP prices the ETF at a discount to NAV to compensate for additional holding costs and mark-to-market costs.

Third, the AP essentially acts as a liquidation agent on behalf of ETF investors. The AP sells the redeemed bonds in a manner that protects remaining ETF and bond investors because she is also a holder of the bonds. Moreover, since the ETF discount is a reflection of the AP's costs, *redeeming* ETF investors ultimately pay these costs through a lower ETF price. Another important observation from our model is that since the AP endogenizes the costs of lower bond prices because of mark-to-market losses, she always avoids fire sales in the bond market. In other words, she acts as a buffer during fire-sale episodes because she endogenously chooses to hold bonds instead of sell them at fire-sale prices. That is, the AP acts as a stabilizing force in the more illiquid bond market because she has "skin in the game." On the other hand, redeeming ETF investors effectively pay high transaction costs when selling ETFs in the secondary market at discounted prices relative to the value

of underlying bonds.

3.3 Creation Model Setup

We now introduce creations, which also helps us illustrate the fundamental differences between premiums and discounts. The creation model set up is the same as that of the redemption model in terms of assets and agents, and the creation basket is also fractional and contains only bond A. Analogous to redemptions, ETF investors exogenously buy $q > 0$ units of the ETF from the AP, and the AP then delivers bonds in the creation basket and exchanges them for ETF shares. We re-interpret γ as the share of bonds that the AP delivers by taking from her own inventory (instead of the share of bonds added to her inventory in redemptions). Thus, $1 - \gamma$ is the fraction of bonds purchased in the bond market. To be consistent with the treatment of holding costs between creation and redemption, we assume that taking bonds from inventory also incurs a cost since the AP deviates from the target inventory (i.e., an opportunity cost). In contrast to redemptions, the price impact cost is now continuous, without a jump in costs at τ: there are no "fire purchases."[19] The representative AP also acts as if driven by perfect competition, which results in setting the *minimum* ETF price (since this is what ETF investors pay).

The AP's payoff for creation is

$$\underbrace{(P_e - NAV_{basket}) \cdot q}_{\text{arbitrage profit}} - \underbrace{\frac{\lambda}{2} \cdot (\gamma q)^2}_{\text{balance sheet cost}} + \underbrace{(z - \gamma q) \cdot \Delta P_A}_{\text{mark-to-market gain}}. \tag{3.5}$$

Fundamental differences between creations and redemptions

Before solving for the creation equilibrium, we emphasize that, in reality, creations are fundamentally different from redemptions for at least two reasons. First, whereas redeemed bonds can always be added to the AP's inventory (irrespective of whether the basket contains bond A or bond B), created bonds cannot always be taken from inventory for the simple

[19] Since the risk of fire sales in times of large ETF redemptions is probably more important for financial stability as described in the introduction, we focus on that case in this paper but the analysis is straightforward to extend for the case of fire purchases. The main conclusions do not change from adding fire purchases.

reason that the AP may not have the specific creation basket bonds in her inventory. In other words, the AP can always add bonds to her inventory but cannot deliver bonds she does not have.[20] Thus, in some cases, the AP *must* buy creation basket bonds. In our model, this means that for some creations, $\gamma = 0$ regardless of the model parameters.

Second, buying creation bonds in the market is beneficial for the AP since it increases the mark-to-market value of her inventory. This means that the mark-to-market term enters as a gain in the AP's payoff for creations since $\Delta P_A > 0$, in contrast to entering as a cost for redemptions (since $\Delta P_A < 0$ in the redemption payoff given in equation 3.1). Thus, the AP might prefer to set $\gamma = 0$ even if she has the creation bonds in her inventory because buying bonds and increasing bond prices raises the mark-to-market value of her inventory. In fact, the AP may even forego arbitrage profits and sell ETF shares at a loss if the benefit from pushing up bond prices outweighs this loss.

3.4 Creation Equilibrium

We now describe the equilibrium holding fraction for creation, γ^*_{CR}, and the resulting premium.

Proposition 2. *(Creation with Fractional Baskets) When the AP delivers bonds to create new ETF shares, the equilibrium share of bonds that are taken from the AP's inventory, γ^*_{CR}, is zero: the AP buys all delivered bonds in the market. The equilibrium ETF price and NAV are*

$$P^*_{e,CR} = P_0 - c(z - q)$$
$$NAV^*_{CR} = P_0 + \frac{0.5 + q}{1 + q} cq,$$

and the equilibrium ETF premium is

$$\pi^*_{CR} \equiv P^*_{e,CR} - NAV^*_{CR} = \frac{0.5}{1 + q} cq - cz. \tag{3.6}$$

[20] It is conceptually possible to borrow bonds to deliver in a creation basket, just as it is possible to borrow bonds to sell short. However, this borrowing involves additional costs and still requires buying the bonds at some point in the future.

The proposition shows that in equilibrium, the AP buys all creation basket bonds in the market, even when she is able to take those bonds from her inventory. This means that the equilibrium choice of the AP is the same regardless of whether she holds the creation basket bonds in inventory or not. While the AP's strategy does not vary depending on her inventory, it still affects the premium.

Remark. The AP always sets $\gamma^*_{CR} = 0$, regardless of her inventory z. However, the ETF premium is different depending on whether the AP has creation basket bonds in inventory, or not. When the AP has no inventory in the basket bonds, the premium π^*_{CR} is always positive. When the AP could use her inventory to deliver bonds but chooses not to do so, the premium can be negative.

The intuition behind creations with a negative premium (discount) is as follows: if the AP's existing inventory in the delivered bonds is very large ($z > \frac{0.5}{1+q}q$), the benefits of boosting the mark-to-market value are also large and the AP is willing to sell the ETF at low prices, resulting in a discount. Creations with a discount go against the classical textbook case where creations always happen at a premium, but are consistent with the evidence in Pan and Zeng (2021). In practice, APs might face additional costs in performing arbitrage (e.g., search costs or creation fees), which would be passed along to ETF traders in the form of higher ETF prices and could shift the premium up enough to make it positive, even when the AP has a large inventory position. For example, subtracting the creation fee times the quantity q from the AP's payoff in equation 3.5 would increase the ETF price and the premium by the fee amount (redemption fees would decrease the ETF price and increase the discount for redemptions).

3.5 Multiple APs and Other Extensions

Our modeling framework is flexible enough to accommodate heterogeneity in initial inventory among multiple APs. This is of particular interest because the assumption of a representative AP may not capture an agile AP with no inventory that opportunistically conducts arbitrage trades. To describe this scenario, we extend the model to allow for two APs, AP_1 and AP_2, with inventory levels z_1 and z_2 where, without loss of generality, $0 \leq z_1 < z_2$. Following the

spirit of a representative AP disciplined by competition, the AP who quotes the best ETF price wins the right to provide liquidity to ETF investors and subsequently redeem or create ETF shares. That is, we denote the AP with the best ETF price as the "marginal" AP. The other aspects of the model are the same and follow the setup described in the baseline model. We formalize the equilibrium for redemptions and creations with multiple APs in the proposition below.[21]

Proposition 3. *(Multiple APs with Heterogeneous Inventory) When there are two APs, AP_1 and AP_2, with inventory levels z_1 and z_2, where $0 \leq z_1 < z_2$, the AP that offers the most attractive price to ETF investors, trades and subsequently creates or redeems ETF shares (i.e., is the marginal AP). For creations, the AP with the largest inventory (AP_2) is marginal. For redemptions, the AP with the smallest inventory (AP_1) is marginal in most cases when $z_1 > 0$, and always when $z_1 = 0$.*

Proposition 3 shows that relaxing the assumption of a representative AP does not change the main conclusion of the model for creation and for redemption if both APs hold some inventory: the marginal AP buys all bonds in creation and holds at least some of the redeemed bonds to protect mark-to-market values in redemption. Redemption with multiple APs does differ from the baseline model if one AP holds no inventory. Such an AP quotes a more competitive price than an AP with positive inventory, does not have an incentive to protect mark-to-market values, and sells all redeemed bonds (as long as it does not trigger a fire-sale). This provides an argument against our inventory-based channel for discounts – even if there are APs that hold inventory, as long as one does not, she does all redemptions.

With multiple APs, if one AP holds no inventory, creations and redemptions are *symmetric* in that the marginal AP always trades all creation/redemption bonds. This is in contrast to our baseline model, which highlights an *asymmetry* between creations and redemptions: the AP trades all creation bonds but only trades a fraction of redeemed bonds (we outline detailed predictions of our baseline model in Section 3.7). We show in Section 4.2 that the

[21] We do not consider strategic play in the case of heterogeneous APs, where one AP can harm the financial condition of other APs by selling bonds that are held by others. In practice, APs hold diversified bond portfolios. While it is probable that there are some bonds that one AP holds but other APs do not, there is very likely significant overlap in bond positions on many dimensions (e.g., issuer and maturity), and reducing bond prices for a few specific bonds is likely to negatively effect at least part of an AP's own holdings.

data clearly support our baseline model and the multiple-AP model where the marginal AP holds inventory. That is, we show that creation and redemption bond volume are not symmetric: APs buy all bonds in creation but sell only a fraction in a redemption.

What might explain why all APs hold at least some inventory? In practice, the AP market is very concentrated with only a few APs that account for the majority of creation/redemption volume (Gorbatikov and Sikorskaya, 2022). Moreover, media reports suggest that one specialized firm is the single largest player in the primary bond ETF market, and accounts for nearly a third of all US Bond ETF secondary-market trading volume. This firm, Jane Street, holds large positions in bonds for several days or weeks, straddling both the algorithmic ETF market and the slower, human-based bond dealer market (Wigglesworth, 2021).[22] This points to the idea that, practically speaking, there are significant barriers to entry: an effective bond AP maintains a modern electronic trading arm for ETFs and a traditional over-the-counter bond market making business. In addition, given creations are much more common than redemptions and the marginal creation AP is the one with the most inventory, inventory-holding APs are likely more present for the typical day-to-day creation event; when a redemption opportunity arises, the same APs may be best positioned to participate. For all of these reasons, and given the empirical evidence in Section 4.2 below, we think it is likely that the APs that typically engage in bond ETF creations and redemptions hold at least some inventory.

We also provide two additional model extensions that address possible asymmetries between the creation and redemption model setups. The first analyzes the possibility that the benefit of temporarily pushing up mark-to-market values (in creations) is smaller than the benefit of preventing temporary declines (in redemptions). This extension to the creation model is presented in Appendix D.1. The second extension allows for disposing of bonds from inventory to be a benefit to the AP instead of a cost of deviating from the AP's target inventory, perhaps by freeing up AP's balance sheet space. We provide this extension in Appendix D.2.

[22] Jane Street is widely recognized as the dominant player in bond ETF markets, in large part because of its role as a top bond dealer. The Federal Reserve confirmed this fact by adding Jane Street to the list of firms that could serve as a counterparty for its bond purchases through the Secondary Market Corporate Credit Facility (SMCCF); most of the listed counterparties are large investment banks (e.g., Citigroup, Bank of America, etc.).

3.6 Mutual Fund Comparison

We now compare the ETF analysis above with a stylized mutual fund (MF). We model a mutual fund as an ETF that uses full baskets (i.e., baskets equivalent to holdings), where the "AP" always sells/buys all bonds to meet redemptions/creations ($\gamma = 0$), and the "AP" has no inventory and only acts as a dealer that trades on behalf of the MF ($z = 0$). We think of the "AP" as the agent that trades on behalf of the mutual fund.[23] We present the formal propositions for this special case of our model in Appendix B. The propositions show that for our stylized mutual fund, the ETF price is equal to NAV, and thus premiums and discounts are always zero.

Our model shows that ETFs are not that different from mutual funds for the case of creations, since the equilibrium γ is zero for both types of funds. However, ETFs and mutual funds are very different when it comes to redemptions. Redemptions of mutual fund shares (outflows) are directly transmitted to the underlying bonds since there is no intermediary to internalize the price impact costs, i.e., there is no buffer between the mutual fund and the bond market. In times of large redemptions (that could trigger a fire sale), the equilibrium mutual fund price is always lower than the ETF price (which can be seen by examining the "Fire Sale Price Impact" line at $\gamma = 0$ vs. the price at $\gamma = \gamma_\tau$ in Figure 3.2). In contrast to mutual fund redemptions, ETF redemptions are not transmitted in full to the underlying bond market because the equilibrium γ is never zero (for an AP with positive inventory in normal times and any AP in times of fire sales). This allows the AP not only to buy ETF shares at a deeper discount from redeeming ETF investors (compared to the mutual fund), but it also protects existing ETF investors and bond markets from the transmission of large selling pressure from the ETF market to the more fragile bond market. In this sense, ETFs are superior to mutual funds in managing illiquid assets, especially in times where large redemptions could trigger a fire sale.

Our model also highlights another important benefit in the design of ETFs: the use of in-kind transfers. When the AP buys the ETF at a discount from selling ETF investors,

[23]The AP for an ETF facilitates inflows and outflows (through creation and redemption). There are no APs for mutual funds. Thus, we use the AP in the model to describe how a mutual fund transacts to handle inflows and outflows.

the AP is effectively buying *bonds* at a discount because she can redeem ETF shares for the underlying bonds. Moreover, because the AP acquires bonds through an in-kind transfer (in the primary ETF market), she receives these bonds *without changing mark-to-market values*. This is not possible with mutual funds. To see why, imagine there is an inventory-holding dealer that trades with a mutual fund facing redemptions. If a mutual fund sells bonds to this dealer to satisfy investor redemptions, either (1) the dealer buys bonds at a discount to fair value and reduces mark-to-market values, or (2) preserves mark-to-market values but buys the bonds without earning a spread. ETF-style in-kind transfers would allow this dealer to decouple the purchase price of bonds from the market price of bonds, and give flexibility to the dealer to act as a buffer.

To avoid fire sales in reality, mutual funds often hold a larger share of liquid assets or cash. This is likely to increase tracking error and create cash drag. We interpret this choice as incorporating the cost of possible liquidation at fire sale prices into other forms, but ultimately the threat of liquidation at fire sale prices is what creates a cost to mutual fund investors in the first place. Mutual funds may also choose to sell more liquid bonds first (akin to the idea of fractional baskets). However, this still leads to secondary market transactions, which in turn affects market prices and mark-to-market values – it does not allow a dealer (or AP-like agent) to receive the bonds in-kind.

The ETF discount from our model could also be useful to calibrate the mutual fund swing factor, which is a discount applied to the price MF investors receive in periods of heavy redemptions. Since APs are also dealers in the bond market, the ETF discount could be interpreted as dealers' costs of handling general fund redemptions, and thus can be used as an estimate of the costs incurred by mutual funds facing redemptions. In that sense, the ETF discount can be used as a real-time, market-based estimate of mutual fund redemption costs, and can thus be applied to calibrate a mutual fund's swing factor, which is typically capped at 2%.[24]

[24]ETF discounts can be much greater than 2%. In fact, they can reach levels that are several times larger than 2% as seen during the March 2020 COVID sell-off, when *average* bond ETF discounts for some bond categories were larger than 6% (see Figure 4.3).

3.7 Model Predictions

Our model generates four testable hypotheses related to bond trading volume, ETF discounts/premiums, AP inventories, and ETF/NAV returns.

Hypothesis 1. *(Creation baskets have greater trading volume than redemption baskets) The AP purchases all basket bonds in response to creation but only sells a fraction of basket bonds in response to redemption.*

Our model highlights the asymmetry between creations and redemptions because of the differential effect of AP's inventory. We test this prediction by looking at trading volume of bonds used in a creation vs. those used in a redemption. Proposition 2 shows that γ^*_{CR} is zero and the AP purchases all bonds needed for creations in the market. That is, for a creation flow q, all of the creation basket bonds are bought in the bond market and delivered to the ETF, i.e., trading volume is equal to q. Thus, the model predicts that creation basket bonds have greater trading volume compared to their non-basket counterparts, reflecting the arbitrage trading of the AP. In contrast, Proposition 1 illustrates that for redemptions, γ^*_{RD} is positive when the AP has some inventory ($z > 0$) and thus the AP puts some fraction of redeemed bonds on balance sheet. For a redemption flow q, only some of the redeemed basket bonds (totaling a value of $(1 - \gamma^*_{RD})\, q < q$), are sold in the bond market, i.e., trading volume is less than q. Thus, the model predicts that redemption basket bonds have more trading volume than the non-basket bonds, but that this volume differential should be smaller than for creation (since the AP trades all created bonds but does not trade all redeemed bonds).

Hypothesis 2. *(Inventories and Discounts/Premiums) For typical redemptions, the equilibrium discount is increasing in the AP's inventory. For creations, the equilibrium premium is decreasing in the AP's inventory.*

An important implication of our model is that it proposes a novel explanation for premiums and discounts based on AP inventories. In the case of typical redemptions (i.e., where selling $1 - \gamma_c$ of the redeemed bonds does not trigger a fire sale), the model shows that AP inventory plays an important role in determining to what extent selling pressure in the ETF is transmitted to the bond market. From Proposition 1, it is simple to show that the

discount is weakly increasing in inventory, i.e., $-d\pi^*_{RD}/dz \geq 0$, and strictly increasing when $\gamma < 1$, i.e., when the AP does not hold all redeemed bonds (see Appendix C for details). That is, greater inventory increases the discount as the AP decides to put a larger fraction of the redeemed bonds in inventory to mitigate mark-to-market losses. The intuition behind this effect is relatively simple. In the case of ETF sales, the AP acts as a buffer between the ETF market and the bond market. With greater inventory, the AP has more skin in the game and sells fewer bonds to protect the mark-to-market value of her inventory, weakening the connection between the two markets. With no inventory, the AP has no skin in the game and sells all redeemed bonds ($\gamma^*_{RD} = 0$) and serves as a pass-through, which reduces bond prices and the ETF discount.

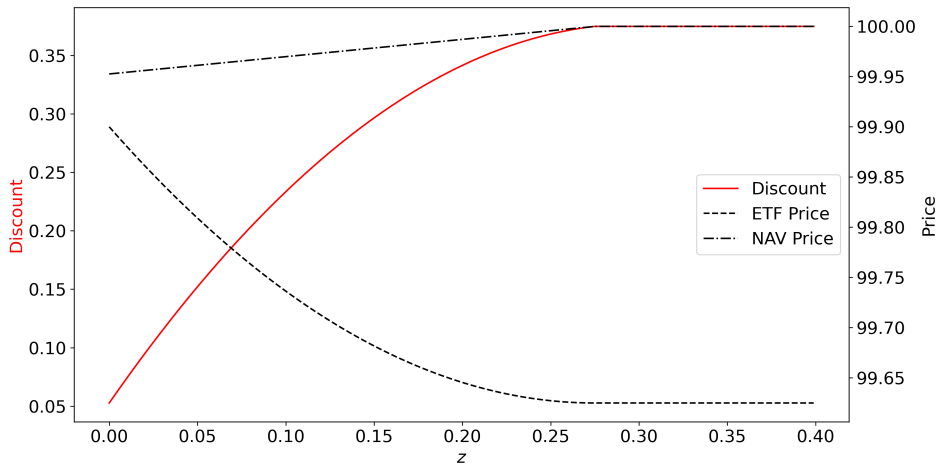
Figure 3.3 provides an example to illustrate these effects. The figure plots the discount (left y-axis) and ETF price and NAV (right y-axis) as a function of the AP's inventory. The discount (solid red line) is increasing in inventory until it hits a point where the AP has such a large inventory ($z > \frac{\lambda - 2c}{c} q$) that she prefers to simply hold all redeemed bonds. After that point, the discount is constant. Additionally, the figure illustrates that the discount almost exclusively stems from a decrease in the ETF price, which reflects greater trading costs for redeeming ETF investors (see Hypothesis 4 below for more on this).

For creations, the model also predicts a negative relation between inventories in premiums. When the AP holds inventory in the created bonds, it is simple to show from Proposition 2 that $d\pi^*_{CR}/dz < 0$, or greater inventory decreases ETF premiums. The intuition behind this case is also rather straightforward: with greater inventory, the AP realizes greater benefits from buying bonds and increasing mark-to-market values, which allows her to reduce the ETF price and premium.

Hypothesis 3. *(Inventories Amplify Discounts During Fire Sales) When ETF redemptions are large enough to trigger a fire sale, the AP's inventory has a greater effect on the discount.*

Hypothesis 3 stems from the model's prediction that the effect of inventory on discounts is greater when ETF redemptions are large enough to potentially trigger fire sales. That is, in the fire sale region described in Section 3.2 when $(1 - \gamma_c)\, q > \tau$, the AP has to hold more bonds in inventory ($\gamma_\tau q$) compared to the non-fire-sale optimum ($\gamma_c q$), where

Figure 3.3: Inventory and Discounts

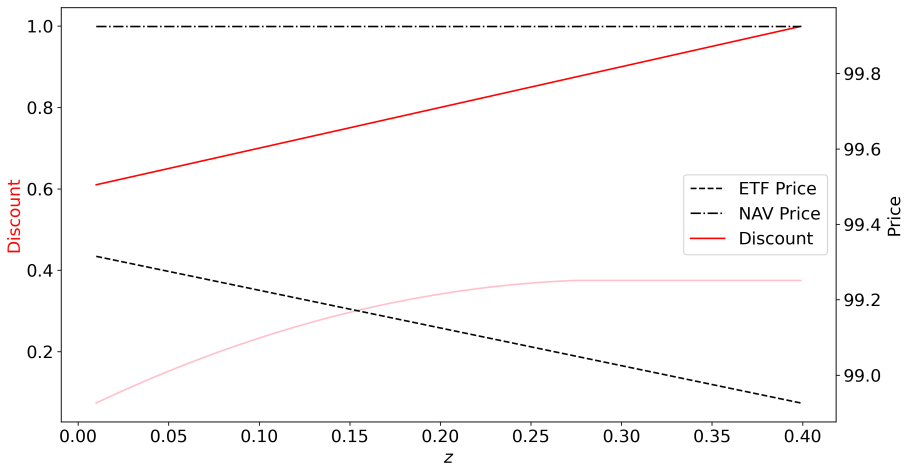


Notes. This figure plots the ETF discount on the left y-axis, and the ETF price and NAV on the right y-axis as a function of the AP's inventory z. The figure describes the case when the redemption quantity q does not trigger a fire sale and the AP chooses $\gamma^*_{RD} = \gamma_c$. For this example, we use the same parameters as in Figure 3.1: $q = 0.05$, $z = 0.2$, $c = 2$, $f = 4$, $\lambda = 15$, $\tau = 0.1$, and $P_0 = 100$. See Section 3.7 for more details on the model predictions, and Section 3.2 for more details on the AP's equilibrium fraction of bonds held.

the additional bonds held are specifically to avoid triggering a fire sale. This additional holding quantity leads to significantly deeper discounts, and these discounts are increasing with greater inventory. In typical redemptions, there is a limit to the effect of inventory on discounts – if inventory is large enough, the AP simply holds all redeemed bonds and the discount is constant (e.g., see Figure 3.3). For fire-sale redemptions, there is essentially no limit to the effect – even very large inventory positions continue to deepen discounts.

Figure 3.4 illustrates these effects. The figure plots the discount (left y-axis) and ETF price and NAV (right y-axis) as a function of the AP's inventory for a large redemption (large q) that has the potential to trigger a fire sale. The figure also shows the discount from Figure 3.3 (small q) in light red to compare the typical and fire-sale discounts. This example unsurprisingly shows that large redemptions lead to much bigger discounts than typical redemptions due to greater costs of handling more redeemed bonds (solid red line vs. light red line). More importantly, the gap between the red line and light red line illustrates that discounts are much larger during fire-sale redemptions, and that the effect is particularly strong for large inventories. Whereas typical redemption discounts are constant after around $z = 0.25$, fire-sale discounts are still increasing after that point, and the gap between the

Figure 3.4: Discounts (Typical vs. Fire Sale)



Notes. This figure plots the ETF discount (solid red line), ETF price (dashed black line), and NAV (dash-dotted black line) as a function of the AP's inventory z for the fire-sale case when the AP chooses $\gamma^*_{RD} = \gamma_\tau$. The figure also plots the discount for a "typical" redemption quantity (solid light-red line), i.e., when the AP does not need to adjust behavior to avoid a fire sale and chooses $\gamma^*_{RD} = \gamma_c$. The discounts are plotted on the left y-axis. The ETF price and NAV are on the right y-axis. For this example, we use the same parameters as in Figure 3.3: $z = 0.2$, $c = 2$, $f = 4$, $\lambda = 15$, $\tau = 0.1$, and $P_0 = 100$; for the typical case we use $q = 0.05$, and for the fire sale case we use $q = 0.20$. See Section 3.7 for more details on the model predictions, and Section 3.2 for more details on the AP's equilibrium fraction of bonds held.

two types of discounts becomes larger. That is, in fire sale cases, greater AP inventory has an increasingly larger effect on the discount compared to typical redemptions.[25]

Hypothesis 4. *(Premiums/Discounts as ETF Trading Costs) Buying (selling) ETF investors pay transaction costs because of temporary price impact in the ETF, which leads to premiums (discounts); ETF prices subsequently reverse.*

The model predicts that the AP's costs of market-making in the ETF (i.e., dealing with creations and redemptions) are a driver of premiums and discounts. Moreover, since our model has no asymmetric information about fundamentals and buyers and sellers in the ETF are uninformed, the ETF price changes can be interpreted as temporary price impact, or trading costs for ETF investors. For example, selling ETF investors trade at relatively low prices (and at a discount relative to the underlying bonds) because APs must generate some revenue to cover the costs associated with market making. While our model is not dynamic in nature, the usual prediction of price impact induced by uninformed trading is

[25]The figure also illustrates that during fire-sale redemptions, almost all of the discount is again driven by a drop in ETF prices and a muted change in NAV.

that it leads to subsequent price reversals (Campbell et al., 1993; Pastor and Stambaugh, 2003). Thus, we should expect both premiums and discounts to correct from subsequent price reversals in the ETF. For example, selling ETF investors sell at relatively low ETF prices and at a discount to NAV, but eventually the discount is eliminated as ETF prices recover from the temporary selling pressure.[26]

Price impact might not be the only explanation for premiums and discounts. Madhavan and Sobczyk (2016) argue that ETF prices reflect price discovery and incorporate information faster than bonds. This interpretation of premiums and discounts implies that price impact in the ETF is permanent as it reflects new information that has not yet been incorporated into bonds, in part because bonds tend to be illiquid and may languish at stale prices. For example, bad news about bonds leads to lower bond ETF prices as traders price in the news, while bond prices are left unchanged due to high trading costs; eventually bond prices decrease to incorporate the news.

These explanations for premiums and discounts have distinct predictions. Our model's *price impact* story says that ETF price changes reflect compensation to the AP and lead to a premium/discount, but eventually these ETF price changes revert. That is, a high premium is related to low future ETF returns. The *price discovery* story of Madhavan and Sobczyk (2016) predicts that ETF price changes reflect new information, while bond prices are stale, but eventually bond prices and NAV catch up to ETF prices as they incorporate the new information with a lag. That is, a high premium is related to high future NAV returns.

The two explanations for premiums and discounts are not mutually exclusive. For example, bad news about bonds could be coupled with liquidity shocks to ETF investors, leading to a decline in ETF prices and a discount from two sources: deteriorating bond fundamentals, and uninformed selling by liquidity-demanding ETF investors. In this scenario, the price discovery and price impact stories could both be true – bond prices could subsequently decline to reflect the bad news discovered in ETF prices, and ETF prices could partially revert, correcting the temporary price impact of ETF investors.

[26]This interpretation of premiums and discounts is similar to the interpretation offered by Cherkes et al. (2009) for closed-end funds, where the premium/discount depends on the benefit of liquidity against the cost of managerial fees.

4 Empirical Support for the Model

We now provide evidence in support of the model's four hypotheses.

4.1 Data

We use ETF Global for data on ETF holdings, flows, shares outstanding and other ETF-level variables. We supplement the ETF data with CRSP, and use CRSP in the case of conflicts between the two datasets. We use Markit for data on bond prices, quotes, and volume (which ultimately come from TRACE). We use data on dealer inventories from the Federal Reserve, which we discuss below in Subsection 4.1.2.

To select our sample of ETFs, we make three sample restrictions. First, we use a sample starting in May 2017 due to a change in the way ETF Global produces their data, which drastically improved the quality and reliability.[27] We end the sample for all of the empirical tests in February 2020 to avoid effects that could reasonably be attributed to actions by the Federal Reserve in the bond and bond ETF markets. We use data from March 2020 in Section 4.4 to separately test the fire-sale predictions of the model. Second, we focus on ETFs that primarily hold (at least 90% of holdings) U.S. corporate bonds with a maturity greater than one year. This criteria eliminates ETFs that hold a significant share of derivatives (which is often the case for leveraged or inverse ETFs), other debt instruments, or even other ETFs. On occasion, we also separate corporate bond ETFs on short and long maturity based on a breakpoint of 5 years, following the inventory data (see below). Third, we only use ETFs that regularly update holdings on a daily basis in order to infer *realized* creation and redemption baskets using the methodology described below.

4.1.1 Inferring Realized Creation/Redemption Baskets

We employ a novel approach to compute realized creation and redemption baskets. We infer baskets based on changes in ETF holdings on days with creation/redemption activity (i.e., days with fund flows). To the best of our knowledge, our paper is the first to employ this approach and analyze realized baskets as opposed to announced baskets (the latter are

[27] ETF Global has reported to us that the methodology to record and deliver data changed in April 2017.

provided by the Depository Trust and Clearing Corporation (DTCC) and utilized in Koont et al. (2022)).

There are important differences between realized and announced baskets. Realized baskets represent the actual portfolio of bonds that were delivered from the AP to the ETF (in a creation) or from the ETF to the AP (in a redemption). Announced baskets are simply an indication of what the ETF would be willing to accept; they do not change much from day-to-day, and often differ considerably from the realized basket. The realized basket is more volatile and involves negotiation between ETFs and APs. For example, one of the largest ETF issuers, BlackRock, uses an internal system that handles custom baskets through which APs propose baskets and ETF managers can accept, modify, or reject these proposals.[28]

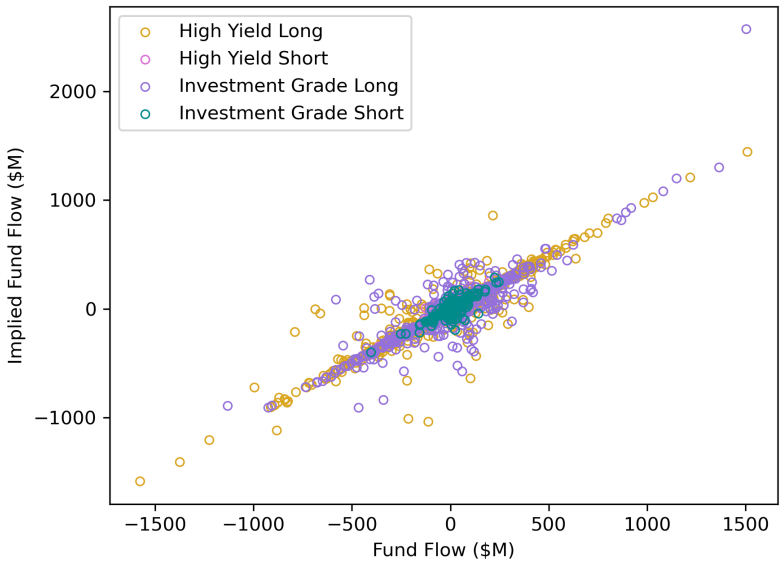
Our methodology implicitly assumes that changes in holdings are mainly driven by creation/redemption activity. Anecdotal evidence from our conversations with several major ETF sponsors and the accuracy of our implied flows (as we show below) suggest that this is a reasonable assumption. Our novel method of inferring creation and redemption baskets is as follows. We calculate the change in number of bonds held for each bond in ETF holdings on each day with non-zero flows. The realized creation basket consists then of bonds with positive holdings changes, whereas the realized redemption basket includes bonds with negative holdings changes. The quantity of each bond in each realized basket is given by the change in daily holdings. We then use bond prices from Markit to compute our measure of implied fund flows, which is the total dollar value of daily bond position changes across all holdings on each day.

The method for inferring baskets requires precise holdings data. The date of holdings in ETF Global is based on the date when ETF issuers report holdings (we have confirmed this with the data provider). We find a systematic data synchronization issue at the ETF sponsor level that require manual adjustment to accurately reflect the correct reporting date of flows and holdings.[29] To account for this, we shift holdings by one day to account for reporting delays that are systematic for some ETF issuers.

[28]BlackRock described this process in a white paper and discussed the process during a webinar marking the public release of the paper (BlackRock, 2021).

[29]Our understanding based on conversations with data providers is that these data synchronization issues stem from different reporting methodologies at the level of ETF issuers. See also Shim and Todorov (2021) for more details.

Figure 4.1: Reported ETF Fund Flows vs. Implied Flows



Notes. This figure provides a scatter plot of reported corporate bond ETF fund flows on the x-axis and our measure of implied fund flows on the y-axis (both in millions of dollars). We use a baseline sample of 89 ETFs and only include observations with non-zero fund flow from May 2017 to March 2020. Our sample has 10,297 observations, and the unit of observation is at the ETF-day level. See Section 4.1.1 for more details on the methodology and data.

In support of our methodology, the median correlation between reported flows and implied flows over all ETFs in our sample is 0.96. A regression of implied flows on reported flows has a slope of 0.94 and an R^2 of 0.85. Figure 4.1 provides a scatter plot of implied flows vs. reported flows for all ETF flow days and confirms the high correlation between the two types of flows.

4.1.2 Inventory Data

We use inventory data from the Federal Reserve Bank of New York, which provides primary dealers' net positions in investment-grade and high-yield corporate bonds for a variety of maturities at a weekly frequency.[30] We aggregate inventories of investment-grade and high-yield bonds between 1 and 5 years (short maturity), and greater than 5 years (long maturity), which are the same breakpoints we use to split the sample of ETFs.[31] The list of primary

[30]The data is provided on the Federal Reserve Bank of New York website (https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics).

[31]The split for inventories does not necessarily match the split for ETFs perfectly, since long maturity ETFs can also hold bonds that we categorize as short maturity.

dealers consists of 25 firms, including the largest, most active bond dealers who are also major APs for the ETFs in our sample. We acknowledge that the inclusion of additional dealers who may not necessarily be APs could add noise to the sample. In order to merge daily ETF premiums to weekly inventories, we match the daily premium with the most recent past inventory position for the corresponding category of bonds that the ETF holds.

One potential concern with our inventory data is that it is aggregated across all dealers and does not reflect the inventory position of the actual AP engaged in creation/redemption. In some sense, aggregate inventory is a better match for our notion of a representative AP, and connects well with the fact that the AP market is extremely concentrated. Data on which AP engaged in a creation/redemption event is not available and thus even data on dealer identities (e.g., from regulatory TRACE) are not very useful. However, given the AP market is concentrated and essentially all participants also serve as bond market makers, we expect inventories to be at least somewhat correlated across dealers at any point in time, especially within a given bond category. To the best of our knowledge, the inventory data we use is the most reliable data on bond dealer inventories that is publicly available.

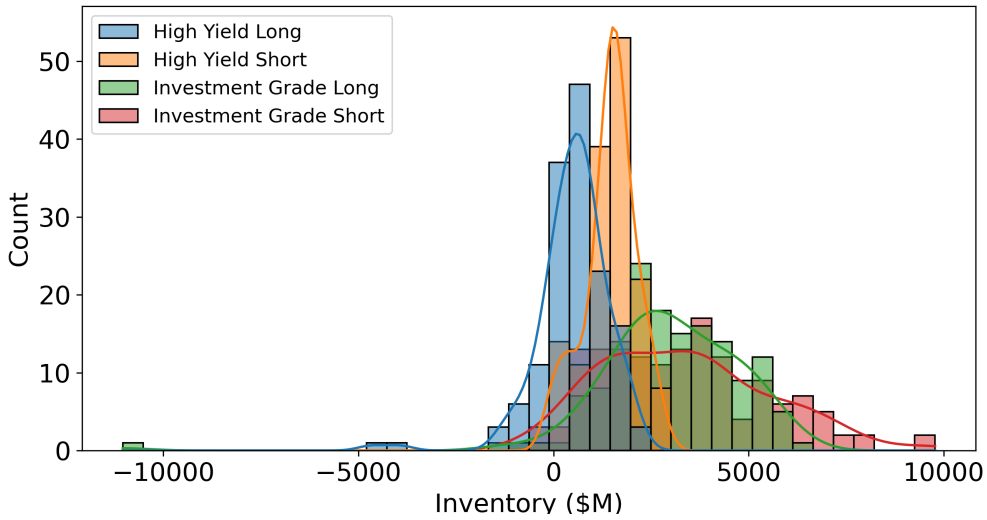
Figure 4.2 presents the distribution of dealer inventories over our sample period for the four corporate bond categories.[32] The figure illustrates two main patterns. First, dealers hold larger inventories in investment-grade bonds than in high-yield bonds. The average inventory position for investment-grade short- and long-maturity bonds is \$3,347M and \$2,998M, respectively. Second, the figure shows that dealers hold relatively little inventory in long maturity high-yield bonds, with an average inventory position of around \$500M. Often, dealers hold no inventory or a net short position in these bonds, as shown in the figure.

4.1.3 Premiums, Liquidity Mismatch, and APs

Table 1 shows that corporate bond ETFs often trade at a premium or discount to NAV and that the magnitude of those price deviations is larger for less liquid corporate bonds compared to more liquid Treasuries. This fact shows that the ETF arbitrage mechanism

[32]In principle, APs in our model could choose to hold the purchased ETF shares instead of redeeming them for bonds, retaining the option to redeem when advantageous. However, the predictions of our model are similar: the only change would be that the inventory held in bonds would be distributed between bonds and ETF shares.

Figure 4.2: Bond Dealer Inventories



Notes. This figure plots the distribution of Federal Reserve primary dealer net positions (i.e., inventories) in corporate bonds as reported by the Federal Reserve Bank of New York. The positions are reported weekly, and we use reported net positions from May 2017 to March 2020 to match the timeline of our ETF sample. To categorize inventories, we use the same maturity cutoffs used for corporate bond ETFs to split the sample into short and long maturity (5 years). See Section 4.1.2 for more details on the data.

works imperfectly for corporate bond ETFs. The typical explanation for premiums and discounts is that they arise because of stale pricing in the illiquid assets underlying the ETF (Petajisto, 2016; Madhavan and Sobczyk, 2016; Lettau and Madhavan, 2018). Recent studies (Shim and Todorov, 2021) show that ETF premiums are a more persistent feature of corporate bond ETFs because of their fractional baskets.

4.2 Trading Volume in Creation vs. Redemption (Hypothesis 1)

In this section, we test Hypothesis 1, which states that basket bonds experience more trading volume for creation compared to redemption. Since the AP acquires all basket bonds as part of creation, we expect that the AP's trading increases the trading volume of basket bonds compared to non-basket bonds by the creation amount. For redemptions, the AP does not trade all basket bonds since she puts at least some of them in inventory. Thus, while redemption basket bonds may also have more trading volume than non-basket bonds, the difference should be smaller than for creations.

Table 1: ETF Premiums and Discounts

ETF Type:	Treasury			Investment Grade		High Yield	
Maturity:	Short	Medium	Long	Short	Long	Short	Long
Mean $\pi_{e,t}$ (bps)	2.70	-4.08	-1.83	10.30	10.41	1.57	6.92
Std. Dev. $\pi_{e,t}$ (bps)	4.42	17.48	33.06	43.71	50.29	47.28	46.73
$\pi_{e,t} > 0$ (% days)	73.99	52.34	60.93	82.11	76.11	63.95	64.85
Mean $\pi_{e,t} \mid \pi_{e,t} > 0$ (bps)	4.36	11.50	11.85	18.13	25.52	24.84	28.07
Mean $\pi_{e,t} \mid \pi_{e,t} < 0$ (bps)	-2.74	-14.56	-22.17	-94.42	-57.26	-39.66	-39.37
Mean bid-ask spread (bps)	1.58	1.97	9.25	15.64	34.96	42.77	47.84

Notes. We use a baseline sample of 111 ETFs from May 2017 to March 2020. For Treasury ETFs, we use three maturity categories: short (ETF's weighted average maturity is less than or equal to 3 years), medium (between 3 and 11 years), and long (greater than 11 years). For corporate bond ETFs, we use only short and long maturity categories, where short maturity are ETFs with a weighted average maturity less than or equal to 5 years, and all other ETFs are long. $\pi_{e,t}$ is the premium for ETF e on day t. In reporting summary statistics, we compute the statistic over all days for each ETF e first, then average over all ETFs in a category. For example, for Std. Dev. $\pi_{e,t}$ for short-maturity investment grade ETFs, we compute the standard deviation of premiums for each ETF in that category, then report the average standard deviation. Mean bid-ask spread is the average bid-ask spread of underlying holdings. See Section 4.1 for more details on the data.

To test this hypothesis, for each day with ETF flows we estimate

$$V_{i,e,t} = \alpha_{e,t} + \alpha_i + \beta \cdot \text{pred\_cr\_vol}_{i,e,t} + \eta \cdot \text{pred\_rd\_vol}_{i,e,t} + \varepsilon_{i,e,t}, \tag{4.1}$$

where $V_{i,e,t}$ is the dollar volume of bond i held by ETF e on a day t with either a net creation or redemption flow. For creation day t, we combine bond volume for each bond i held by ETF e from the day of creation and the day *before*; for redemption, we combine volume from the day of redemption and the day *after*. We do so to account for slight differences in the timing of AP trading and ETF flows. For example, a creation basket may be acquired by buying bonds the day before and the day of ETF creation, whereas redemption bonds could be sold on the day of redemption and the day after it.[33]

The independent variables, $\text{pred\_cr\_vol}_{i,e,t}$ and $\text{pred\_rd\_vol}_{i,e,t}$, are the predicted trading volume for bonds in the creation and redemption baskets, respectively. We compute $\text{pred\_cr\_vol}_{i,e,t}$ ($\text{pred\_rd\_vol}_{i,e,t}$) by computing what the trading volume would be for each bond i if the AP were to buy all of the delivered (sell all of the redeemed) basket bonds in the bond market. For example, if there is a creation of \$10M, this means that \$10M worth

[33]We find qualitatively similar results using only volume on the day of the flow, as well as when we combine volume on the day before, the day of, and the day after creation/redemption.

of bonds are delivered to the ETF. If bond i's weight in the creation basket is 0.1, this means that \$1M worth of bond i is delivered to the ETF. Thus, if the AP purchases all bonds in the creation basket to deliver to the ETF, then the predicted volume in bond i is \$1M. In equations,

$$\text{pred\_cr\_vol}_{i,e,t} = \text{cr\_flow}_{e,t} \times \text{cr\_basket\_weight}_{i,e,t}$$
$$\text{pred\_rd\_vol}_{i,e,t} = \text{rd\_flow}_{e,t} \times \text{rd\_basket\_weight}_{i,e,t},$$

where the flow variables are the magnitudes of the flow in ETF e on date t, and the weight variables are the basket weights for a bond i in a creation or redemption basket of ETF e on date t. If a bond is held by the ETF but is not included in the basket, we set the predicted volume to 0.

Since the purpose of the regression is to analyze the excess volume in basket bonds on flow days, we compare basket bonds with other bonds held by the ETF but not included in the basket (non-basket bonds). To do this comparison, we include *ETF-by-date* and bond fixed effects, similar to Reilly (2021). By including these high-dimensional fixed effects, we can isolate variation *within an ETF on a particular day*, and control for time-invariant bond characteristics. We also control for each bond's liquidity by including the bond's bid-ask spread and for the relative size of a bond in an ETF by including the bond's weight in the ETF. In addition, since bond trading volume tends to be somewhat persistent, we also include the bond's trading volume over the past week (excluding the most recent day). We cluster standard errors at the bond level and the ETF-by-date level.

Our model predicts that for creation, all bonds are purchased in the market (i.e., no bonds are delivered from the AP's inventory, or $\gamma^*_{CR} = 0$). This corresponds to an estimate of $\beta = 1$ in equation 4.1. That is, the excess volume of creation basket bonds over non-basket bonds for an ETF-date corresponds one-for-one with the trading volume from the AP's purchase of each creation basket bond in the bond market. For redemption, the model predicts that the AP trades a fraction of the bonds in the market. This corresponds to an estimate of $0 \leq \eta < 1$ in equation 4.1. That is, the excess volume of redemption basket bonds over non-basket bonds is less than what is predicted if the AP were to sell all the

redemption basket bonds in the bond market.

Another interpretation of the regression estimates is that they allow us to identify γ^*_{CR} and γ^*_{RD} in the data. Our model shows that $\gamma^*_{CR} = 0$ (or $1 - \gamma^*_{CR} = 1$), which means that all creation bonds are purchased in the market. Since $\text{pred\_cr\_vol}_{i,e,t}$ is equal to the expected excess volume if the AP purchases all creation bonds, $\gamma^*_{CR} = 0$ corresponds to $\beta = 1$. Similarly, $\text{pred\_rd\_vol}_{i,e,t}$ is the expected excess volume if $\gamma^*_{RD} = 0$, which would correspond to $\eta = 1$. Since our model predicts $0 < \gamma^*_{RD} \leq 1$ as long as the AP has at least some inventory, η is an estimate of the fraction of redeemed bonds that are traded $(1 - \gamma^*_{RD})$.

One concern is that a bond may be included in many ETF baskets on the same date, especially for redemptions. This would bias our estimate of β and η towards zero since the predicted volume for an ETF-date does not capture additional predicted volume for the same bond from other ETFs on the same date. To account for bonds included in baskets by multiple ETFs on a given date t, we estimate a second regression, which combines the predicted volume for a bond across all ETFs with creation or redemption on that date.[34] Specifically, we set $\text{pred\_cr\_vol}_{i,t} = \sum_e \text{pred\_cr\_vol}_{i,e,t}$ and $\text{pred\_rd\_vol}_{i,t} = \sum_e \text{pred\_rd\_vol}_{i,e,t}$, and estimate

$$V_{i,t} = \alpha_t + \alpha_i + \beta \cdot \text{pred\_cr\_vol}_{i,t} + \eta \cdot \text{pred\_rd\_vol}_{i,t} + \varepsilon_{i,t}.$$

Since the regression is estimated at the bond-date level instead of the bond-ETF-date level like before, we include only bond and date fixed effects, and cluster standard errors at the bond and date level. The model's prediction for this regression is just as above: $\beta = 1$ and $0 \leq \eta < 1$.

Table 2 provides the regression estimates for the two regressions described above, and supports the predictions of the model. The table shows that estimates of β (coefficient on $\text{pred\_cr\_vol}_{i,e,t}$ and $\text{pred\_cr\_vol}_{i,t}$) are indeed close to 1 in all specifications. These estimates provide empirical support to the model's prediction that the AP purchases all creation bonds in the market: the additional volume for basket bonds over non-basket bonds corresponds almost one-for-one with the predicted AP's trading volume. Moreover, we find

[34] We only consider the cases when a bond is included in many creation baskets or many redemption baskets. If a bond is included in both creation and redemption baskets on the same date, we exclude it from the regression as these instances may capture cases when the AP simply moves bonds from one ETF to another with no trading required.

Table 2: Basket Trading Volume

Dependent Variable:	(1) $V_{i,e,t}$	(2) $V_{i,e,t}$	(3) $V_{i,t}$	(4) $V_{i,t}$
pred_vol_cr$_{i,e,t}$	1.159***	1.053***		
	(8.041)	(7.682)		
pred_vol_rd$_{i,e,t}$	0.160***	0.145***		
	(2.855)	(2.875)		
pred_vol_cr$_{i,t}$			1.122***	0.903***
			(7.471)	(7.450)
pred_vol_rd$_{i,t}$			0.205***	0.173***
			(2.857)	(2.856)
ETF×Date F.E.	✓	✓		
Date F.E			✓	✓
Controls		✓		✓
R^2	0.410	0.423	0.418	0.499
Observations	9,195,679	9,195,679	4,207,643	4,207,643

Notes. This table presents estimates from regressing trading volume for bonds held by ETFs on flow days, on creation/redemption variables. For specifications (1) and (2), we regress bond volume on the predicted volume if an AP buys/sells all creation/redemption bonds in the bond market. For specifications (3) and (4), we repeat (1) and (2) but aggregate predicted volume across ETFs for each bond-date. All specifications include bond fixed effects and volumes are measured in $ billion. Specifications (1)-(2) are estimated also including high-dimensional ETF×date fixed effects, which allow us to compare trading volume for bonds included in a basket against other bonds held by the ETF on the same day but not included in the basket. Specifications (3)-(4) use predicted volume aggregated across ETFs on each date, and thus are estimated including date fixed effects (since there is no ETF-specific variation). The dependent variable includes trading volume on the day of the flow and the day before (after) for creation (redemption) to allow for slight differences in the timing of ETF flows and trading in the underlying bonds. In some specifications, we include additional controls: the bond's bid-ask spread, weight in the ETF (or average weight across ETFs), and trading volume over the previous 5 days before creation/redemption. The sample runs from May 2017 to February 2020, and standard errors are clustered at the bond and ETF×date level for (1)-(2) and at the bond and date level for (3)-(4). One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively, and t-stats are shown in parentheses. See Section 4.2 for more details.

that the additional volume for redemption basket bonds over non-basket bonds is between 0 and 1, consistent with the model. The regressions also provide an estimate of the fraction of redeemed bonds traded η, or $1 - \gamma^*_{RD}$ in the model, of around 0.15 to 0.2. These estimates suggest that about 80% to 85% of the redeemed bonds are held by APs and only 15% to 20% are sold in the market.[35]

Through the lens of the model, these estimates for γ^*_{RD} suggest that APs may hold significant inventory and would prefer to hold a rather large proportion of the redeemed bonds to protect mark-to-market values, while buying the ETF at a discount. These results are consistent with media reports quoting ETF APs. For example, Johnson (2023) provides support to the mark-to-market channel of our model by quoting one executive at ETF market maker, GHCO, who said "If the bonds don't trade then you don't have to write them down." The article goes on to state that GHCO bought bond ETFs at as much as a 20 percent discount during the COVID crash and held the assets, leading to significant profits.

These findings are consistent with Hypothesis 1 and the baseline model where the AP holds at least some inventory. This suggests that holding costs are not prohibitively high ($\lambda < \bar{\lambda}$) and that if there are heterogeneous APs, each holds at least some inventory (see Section 3.5 for more discussion).

4.3 Premiums/Discounts and Inventory (Hypothesis 2)

In this subsection, we test Hypothesis 2: our model predicts a positive relationship between inventory and discounts (a negative relationship between inventory levels and premiums), even in non-fire sale episodes. We test this relationship by regressing ETF discounts on inventory in a panel setting for days with redemptions, or

$$-\pi_{e,t} = \alpha_e + \beta \cdot z_{e,t} + x'_{i,t}\delta + \varepsilon_{e,t}, \tag{4.2}$$

where $-\pi_{e,t}$ is the discount (negative of the premium) for ETF e on day t. $z_{e,t}$ is the primary dealer inventory in the bond category corresponding to ETF e (e.g., if e is a long-maturity investment-grade ETF, $z_{e,t}$ is the primary dealer inventory in long-maturity investment grade

[35] We also provide an additional regression specification using standardized volumes in Appendix E.1.

bonds). In some specifications, we also add a vector of control variables, $x'_{i,t}$, which include the signed ETF flow and the basket fraction since these variables are also related to premiums/discounts (Shim and Todorov, 2021). We estimate a similar regression for days with creations:

$$\pi_{e,t} = \alpha_e + \beta \cdot z_{e,t} + x'_{i,t}\delta + \varepsilon_{e,t},$$

where $\pi_{e,t}$ is the premium. These two regressions mimic the two main settings of our model. In all specifications, we also include ETF fixed effects to control for differences between ETFs with respect to creation/redemption policies, which may impact the timing of creations and redemptions as well as the fraction of bonds selected in baskets. We also separately estimate the regressions with ETF issuer fixed effects instead of ETF fixed effects, since creation/redemption policies may be set at the issuer level. We cluster standard errors by ETF and date.

The regression results in Table 3 are consistent with Hypothesis 2, both for the specification with ETF fixed effects (Panel (a)) and issuer fixed effects (Panel (b)). For redemptions (columns 1–3), there is a statistically significant positive relationship between inventory and discounts: a billion dollar increase in inventories raises discounts by about 1 basis point. In the case of creations (columns 4–6), there is a negative and statistically significant relationship between inventory and premiums. These findings support one of the model's main insights: greater inventories mean that the AP holds more redeemed bonds and passes additional holding costs to ETF investors through deeper discounts.

One concern is that inventories and discounts could be correlated for reasons other than APs' actions. For example, the correlation between inventories and discounts/premiums could be jointly driven by the illiquidity of ETF holdings. To address this concern, we include the bid-ask spread of ETF holdings – as bid-ask spreads widen, dealers may adjust their optimal inventory levels and discounts/premiums may be affected by greater costs of ETF arbitrage. Another factor influencing both discounts/premiums and inventories is market stress, when arbitrage spreads typically widen. If dealers' willingness to take on inventory also changes during those times, discounts/premiums would co-move with inventories. To address this concern, we include VIX (as a proxy for market stress) and the return on an

aggregate bond market index. The regression estimates including these additional control variables are provided in specifications (3) and (6). We find that the results are largely unchanged, except for a small drop in the magnitude and statistical significance in some specifications.

As mentioned in Section 4.1.2, we fully acknowledge that the inventory data is far from perfect. In particular, aggregating inventories across dealers omits potentially important variation. This aggregation is likely to work against finding empirical evidence in support of our model since it introduces noise because it adds inventory from dealers that are likely not APs. We find it reassuring that despite this concern, we are able to find a relatively robust relationship between inventories and discounts/premiums, especially after accounting for more standard explanations like liquidity and general market stress. While the empirical evidence in Table 3 is far from causal, these facts combined with the channel uncovered by the model support our paper's overarching story about the role that inventory plays in determining discounts and premiums.

4.4 Discounts and Inventory during COVID (Hypothesis 3)

In this subsection, we test Hypothesis 3: inventory has a greater effect on discounts during periods of large redemptions. The COVID period provides a useful setting to test the fire-sale implications of our model.[36] Because of the considerable amount of bond ETF selling and redemptions during March 2020, APs likely faced the trade-off articulated in our model: sell bonds at fire-sale prices or hold bonds in inventory to reduce mark-to-market losses.

Figure 4.2 shows that dealers hold the largest inventory in short-maturity investment-grade bonds, followed by long-maturity investment-grade bonds. The data also illustrates that dealers typically hold relatively little to no inventory in long-maturity high-yield bonds. Based on these patterns, our model predicts that discounts should be deeper for investment-grade ETFs compared to high-yield ETFs, and especially so for short-maturity investment-

[36]There are two other reasons to separate the COVID episode and to analyze it outside the baseline regression framework. First, discounts were so large for many ETFs during the COVID period that they essentially become outliers, making regression analysis less suitable. Second, the COVID period economically represents a single "event" instead of multiple days of independent discount observations. Moreover, the event spans 10-20 days but, given the weekly frequency of the inventory data, amounts to only a few observations. For these reasons, a regression analysis is less useful.

Table 3: Premiums and Inventory

(a) ETF Fixed Effects

Dependent Var:	ETF Discount (bps)			ETF Premium (bps)		
	(1)	(2)	(3)	(4)	(5)	(6)
Inventory ($B)	1.107***	1.019***	0.705*	-0.493**	-0.463**	-0.573***
	(2.627)	(2.687)	(1.777)	(-2.123)	(-1.998)	(-2.642)
Fund Flow ($M)		-0.025***	-0.016***		0.043**	0.046**
		(-5.258)	(-3.246)		(2.412)	(2.602)
Basket Percentage (%)		-0.021	-0.023		-0.068*	-0.070**
		(-0.541)	(-0.669)		(-1.796)	(-2.070)
Bid-Ask Spread (bps)			0.304**			-0.179
			(2.461)			(-1.553)
VIX			1.451***			-1.121***
			(4.365)			(-6.935)
Aggregate Bond Ret. (bps)			-0.018			0.018
			(-0.411)			(0.851)
R^2	0.333	0.346	0.441	0.203	0.208	0.228
Observations	2,370	2,370	2,370	6,378	6,378	6,378
Sample	Redemption Only			Creation Only		

(b) Issuer Fixed Effects

Dependent Var:	ETF Discount (bps)			ETF Premium (bps)		
	(1)	(2)	(3)	(3)	(5)	(6)
Inventory ($B)	0.850**	0.799**	0.623**	-0.957*	-0.824*	-0.776*
	(2.320)	(2.232)	(1.970)	(-1.893)	(-1.772)	(-1.931)
Fund Flow ($M)		-0.021**	-0.014**		0.011	0.008
		(-3.670)	(-2.461)		(0.751)	(0.597)
Basket Percentage (%)		-0.055*	-0.052*		0.095	0.101
		(-1.840)	(-1.920)		(1.491)	(1.535)
Bid-Ask Spread (bps)			0.049			0.073
			(0.819)			(0.852)
VIX			1.413***			-1.024***
			(4.426)			(-6.102)
Aggregate Bond Ret. (bps)			-0.034			0.021
			(-0.781)			(0.785)
R^2	0.290	0.302	0.389	0.114	0.119	0.137
Observations	2,385	2,385	2,385	6,381	6,381	6,381
Sample	Redemption Only			Creation Only		

Notes. This table estimates panel regressions of discounts (specification 1-3) and premiums (4-6) on inventory levels with various controls: fund flows, basket percentage, bid-ask spread, VIX, and returns on an aggregate bond index. Panel (a) presents the estimates including ETF fixed effects (F.E.) and panel (b) uses issuer F.E. Specifications (1) to (3) use only redemption days and specifications (4) to (6) use only creation days. Since inventories are reported weekly, we use the most recently reported past inventory level for each bond category. Standard errors are clustered by ETF and date and t-stats are shown in parentheses. We use the corporate bond ETFs from our baseline sample of ETFs from May 2017 to February 2020. We focus on the pre-COVID period due to extremely large (negative) premium outliers during the COVID market sell-off. We analyze the COVID period separately in Section 4.4. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section 4.1 for more details on the data.

grade ETFs, where AP inventories are the largest.[37]

This prediction goes against conventional wisdom. As articulated in Hypothesis 4, one view of bond ETF discounts during the COVID crisis is that they arose because ETFs were "leading the market" and serving as a vehicle for price discovery, while bond prices were stale and left behind (Madhavan and Sobczyk, 2016). This story implies that the illiquidity of bonds is the cause of discounts during this period of market stress: the high trading costs of illiquid bonds disincentivize traders from selling bonds to match prices discovered in the ETF. Quite intuitively, this view then predicts that ETFs holding the most illiquid bonds should trade at the steepest discounts since prices are more stale for more illiquid instruments. That is, this story predicts that high-yield bonds and longer-maturity bonds should have the greatest discounts, which is, coincidentally, the opposite of our model's prediction.[38]

Figure 4.3 shows the pattern of normalized ETF prices in Panel (a) and average ETF premiums in Panel (b) for four bond ETF categories during the COVID market sell-off in March 2020. We compute normalized ETF prices by dividing each ETF's price over this period by its price on February 21st (before the sell-off commenced), and then take the AUM-weighted average of these normalized prices for each day in a particular bond ETF category. We also compute the AUM-weighted average premium over all ETFs within a bond ETF category per day. Panel (a) shows that the ETFs holding the most illiquid assets had the largest price declines, with long-maturity and short-maturity high-yield ETFs exhibiting the largest decline, followed by long-maturity investment-grade ETFs; short-maturity investment-grade ETFs had a noticeably shallower decline in prices.

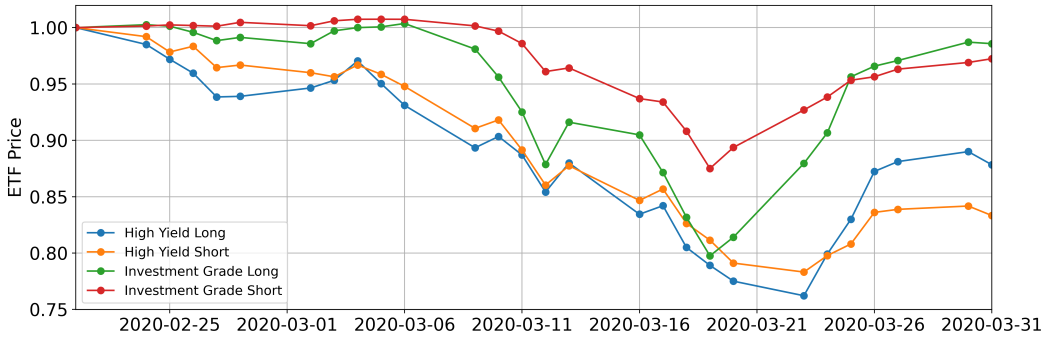
Panel (b) paints a very different picture for discounts compared to prices. While price declines were steepest in ETFs holding the most illiquid assets, discounts were steepest for ETFs holding bonds where APs hold greater inventory. The figure shows that short-maturity

[37]The weekly timing of the inventory data from the Federal Reserve makes direct comparisons between inventory and premiums challenging. In particular, the deepest discounts occurred between Wednesday March 18, 2020 and Friday March 20, 2020. On Monday March 23, 2020, the Federal Reserve announced the Secondary Market Corporate Credit Facility (SMCCF) with plans to purchase investment-grade bonds and bond ETFs, triggering a complete reversal in ETF discounts. But since the Federal Reserve provides inventories as of each Wednesday, we do not observe how inventories adjusted in response to deep discounts or the SMCCF individually, but only the combined change.

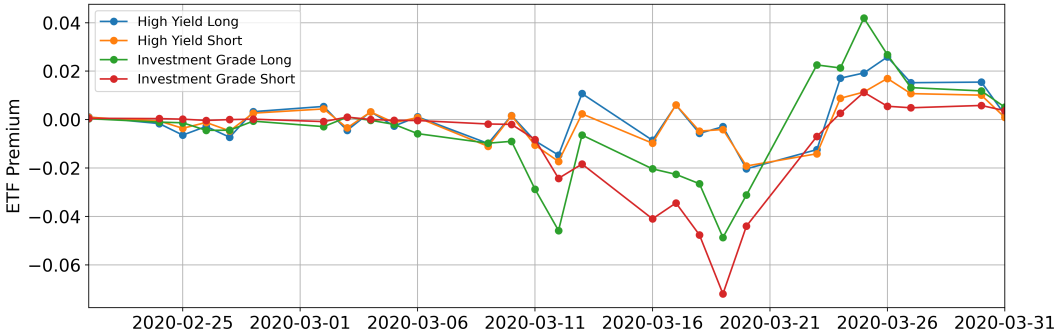
[38]Table 1, which provides average bid-ask spreads by bond type, supports this ordering of illiquidity.

Figure 4.3: Bond ETF Prices and Premiums during the COVID Episode

(a) ETF Prices



(b) ETF Premiums



Notes. Panel (a) shows the normalized ETF prices and Panel (b) provides the average premium by ETF category around March 2020. ETF prices are normalized to 1 at the beginning of the sample and we take the AUM-weighted average price for each trading day within an ETF category. We also take the AUM-weighted average of premiums for each trading day and ETF category. We use the corporate bond ETFs from our baseline sample of ETFs. See Section 2 for more details on the data.

investment-grade ETFs, i.e., those holding the *most liquid* corporate bonds, experienced the deepest discounts. Long-maturity investment-grade ETFs also saw large discounts, whereas high-yield ETFs experienced the smallest discounts. In fact, the discounts in short-maturity investment-grade ETFs were between 2 and 6 times larger than those for high-yield ETFs. These findings align well with the prediction from our model but go contrary to the conventional price discovery view that premiums and discounts are largely driven by illiquidity, which prevents APs from fully exploiting ETF arbitrage opportunities.

Our model's inventory-based explanation for ETF discounts helps understand why ETFs holding more *liquid* bonds saw deeper discounts during the March 2020 sell-off. While Figure 4.3 is consistent with the model's story, it is a simple illustration that comes from a single disruptive shock. That being said, alternative stories fall short in explaining the puzzling dynamics of ETF discounts during COVID because they either (1) predict an asymmetric effect that leads to high-yield ETFs having *larger* discounts than investment grade ETFs (e.g., stale prices due to illiquidity, larger holding costs for high-yield bonds) or (2) cannot explain an asymmetric effect between high-yield and investment-grade ETFs (e.g., general increase in volatility, limited arbitrage capital).

Our results illustrate that ETFs can keep liquidity shocks at least somewhat isolated within the ETF market and may not transmit them in full to the more fragile bond market. This key result of our paper relies on the AP holding some inventory, because then the AP has skin in the game; i.e., after redeeming bonds, she refrains from trading. The intuition from our model squares well with the comments of an executive from one bond ETF market-making firm, who stated

> "The traditional approach across the entire fixed-income industry to a bond crisis is to do nothing. The default behaviour is that everyone — *outside the ETF creates and redeems* — just turns their chair, looks away and says 'we will wait it out.' [emphasis added]" (Johnson, 2023)

4.5 Premiums/Discounts as Trading Costs (Hypothesis 4)

In this section, we test whether premiums and discounts represent ETF trading costs, as suggested by our model and outlined in Hypothesis 4.

The model shows that the AP clears the market by trading with ETF investors, and sets the ETF price based on the costs of handling creation/redemption bonds. If the costs are high, this leads to low (high) ETF prices and larger discounts (premiums) relative to NAV. This is the sense in which ETF premiums and discounts represent trading costs for ETF buyers and sellers. If this view is correct, then ETF prices should revert as temporary price pressure from uninformed ETF investors subsides.

To test this hypothesis, we examine cumulative ETF and NAV returns over 5 trading days following a creation or redemption.[39] We separately analyze a sample of all creation days with a premium and all redemption days with a discount using our pre-COVID dataset. For each sample, we regress 5-day ETF and NAV returns following a creation or redemption on the ETF premium/discount, while controlling for the ETF return and NAV return on the creation/redemption day, or

$$r^{etf}_{e,t\to t+5} = \alpha_e + \alpha_t + \beta_{etf}\pi_{e,t} + \gamma_{etf} r^{etf}_{e,t} + \eta_{etf} r^{nav}_{e,t} + \varepsilon^{etf}_{e,t},$$
$$r^{nav}_{e,t\to t+5} = \alpha_e + \alpha_t + \beta_{nav}\pi_{e,t} + \gamma_{nav} r^{etf}_{e,t} + \eta_{nav} r^{nav}_{e,t} + \varepsilon^{nav}_{e,t},$$

where $r^{etf}_{e,t\to t+5}$ and $r^{nav}_{e,t\to t+5}$ are the cumulative ETF return and NAV return for ETF e over the 5 days following the creation/redemption day t, $\pi_{e,t}$ is the ETF premium, $r^{etf}_{e,t}$ and $r^{nav}_{e,t}$ are the ETF and NAV return on day t, and α_e and α_t are ETF and day fixed effects. We cluster standard errors at the date and ETF level. Note that here we use premiums ($\pi_{e,t}$) for both the creation *and* redemption samples, instead of premiums for creation ($\pi_{e,t}$) and discounts for redemption ($-\pi_{e,t}$) as in Table 3 for simpler interpretation of the regression coefficients. For example, for the sample where we examine redemption days with a discount, the premium $\pi_{e,t,}$ is negative.

Our model's trading cost interpretation of premiums and discounts predicts that $\beta_{etf} < 0$,

[39] We reach qualitatively similar conclusions by examining a shorter period of returns (only the next trading day) or a longer period (the next 10 or 20 trading days).

which means that high premiums forecast low ETF returns, as ETF prices revert (similarly, low negative premiums forecast high ETF returns). The model also predicts that β_{nav} should be weakly negative, but its magnitude should be smaller than that of β_{etf}. The reason is that the AP also trades bonds in response to flows (more for creation than redemption) and hence, some of the price impact in the ETF market spills over to the underlying bonds and NAV as in Ben-David et al. (2018).

The price discovery view, outlined in Hypothesis 4, argues that ETF prices reflect information quickly, while the underlying assets' prices are stale due to low trading volumes and illiquidity. Thus, the price discovery view suggests that the NAV should "catch up" with ETF price changes over subsequent periods. This alternative view predicts then that $\beta_{etf} = 0$ and $\beta_{nav} > 0$, i.e., high premiums predict high NAV returns (low negative premiums predict low NAV returns).[40]

Table 4 provides the regression estimates, and shows that the data only supports our model's trading cost interpretation of premiums and discounts (Hypothesis 4). First, the table shows that $\beta_{etf} < 0$ for premiums (Panel (a)) and discounts (Panel (b)), suggesting that premiums and discounts are corrected through significant reversal in ETF prices. Moreover, the magnitudes are large and range from around -0.4 to -0.8; as a benchmark, a coefficient of -1 means that over the following 5 days, the ETF return reverses by exactly the magnitude of the discount. Second, when describing Hypothesis 4 above, we also raised the possibility that both the trading costs story and the price discovery story could be true in the data. However, the data do not support the price discovery story: the table shows that β_{nav} is small in magnitude and statistically insignificant. This shows that, on days with flows, ETF premiums and discounts do not subsequently lead to NAV "catching up."

In fact, the regression estimates provide new insight into ETFs and price discovery. The results point to two possibilities (not mutually exclusive): for days with ETF flows, either price discovery does not lead to premiums or discounts, or instances of price discovery occur infrequently enough relative to uninformed trading and do not show up in estimates of β_{nav}. The former interpretation implies that when price discovery occurs in bond ETFs, it also

[40]We focus on creation days with a premium and redemption days with a discount because the price discovery story assumes the ETF "leads" the market and thus, creation (redemption) days with a discount (premium) are inconsistent with that hypothesis.

Table 4: Premiums/Discounts and Returns

(a) Premiums

	(1)	(2)	(3)	(4)
Dependent Variable:	$r^{etf}_{e,t\to t+5}$	$r^{etf}_{e,t\to t+5}$	$r^{nav}_{e,t\to t+5}$	$r^{nav}_{e,t\to t+5}$
$\pi_{e,t}$	-0.759***	-0.554***	0.058	0.094
	(-13.044)	(-7.793)	(1.532)	(1.387)
$r^{etf}_{e,t}$		-0.139		0.074
		(-1.629)		(0.879)
$r^{nav}_{e,t}$		0.274***		0.060
		(3.097)		(0.693)
R^2	0.474	0.474	0.399	0.401
Observations	6,110	6,110	6,110	6,110

(b) Discounts

	(5)	(6)	(7)	(8)
Dependent Variable:	$r^{etf}_{e,t\to t+5}$	$r^{etf}_{e,t\to t+5}$	$r^{nav}_{e,t\to t+5}$	$r^{nav}_{e,t\to t+5}$
$\pi_{e,t}$	-0.420*	-0.626**	0.204	-0.089
	(-1.901)	(-2.404)	(1.142)	(-0.358)
$r^{etf}_{e,t}$		0.329*		0.470**
		(1.708)		(2.059)
$r^{nav}_{e,t}$		-0.050		-0.095
		(-0.283)		(-0.521)
R^2	0.680	0.685	0.644	0.655
Observations	954	954	954	954

Notes. This table presents estimates from regressing cumulative ETF returns $r^{etf}_{e,t\to t+5}$ and NAV returns $r^{nav}_{e,t\to t+5}$ over the 5 days following an ETF flow on the ETF premium, as well as ETF and NAV returns on the day of the flow t. We present the estimates separately for creation days with a positive premium (Panel (a)) and redemption days with a discount (Panel (b)). All regressions also include date and ETF fixed effects. Returns and premiums are in basis points. The sample runs from May 2017 to February 2020, standard errors are clustered at the ETF and date levels, and t-stats are shown in parentheses. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section 4.5 for more details.

occurs in the underlying bonds. The latter interpretation implies that most premiums and discounts are driven by noise traders instead of informed traders.

5 Conclusion

We develop a model to understand how APs transmit shocks from corporate bond ETFs to the underlying illiquid assets. The AP acts as an arbitrageur between the ETF and the bond market, and provides liquidity to uninformed ETF traders. The AP sets the price at which she trades with ETF investors by incorporating the costs of handling illiquid bonds.

Our model provides several new insights. First, the AP behaves asymmetrically: she transmits all buying from the ETF to the underlying bonds but only a fraction of the selling. Second, inventory gives the AP "skin in the game" and makes her less willing to sell redeemed bonds because the AP protects the mark-to-market value of her inventory. Third, the AP always avoids a fire sale since she internalizes the cost of lower bond prices. Fourth, the AP passes holding costs to selling ETF investors in the form of larger discounts.

Our paper's findings provide an important insight into investment vehicles managing illiquid assets. We show that a stylized mutual fund – one that sells assets to handle redemptions – may trigger fire sales in illiquid assets in times of large redemptions. In response, mutual funds may hold cash or liquid assets as a precaution, leading to tracking error and cash drag. In contrast to a mutual fund, an ETF relies on an AP to handle the orderly liquidation of illiquid assets, and our model shows that in equilibrium the AP holds bonds to avoid triggering a fire sale. Since an ETF hands off redemptions to an AP who already holds the underlying assets, the AP acts as an agent on behalf of bond investors and remaining ETF investors, and redeeming investors bear the costs of preventing a fire sale in the form of deep discounts. In this sense, our paper shows that ETFs may be a superior investment vehicle over mutual funds in managing illiquid assets.

In this light, our paper also has important implications for policy makers, who have raised concerns that ETFs may present a systemic risk if they trigger the rapid selling of illiquid assets. Our paper helps understand how ETFs may or may not present a risk to the functioning of underlying illiquid asset markets. When APs hold inventory, ETFs mitigate

the risk of fire sales because APs act as a buffer between the ETF and the underlying assets, while still providing liquidity to redeeming ETF investors, albeit at potentially large discounts. When APs do not hold inventory, ETFs may propagate shocks to illiquid markets because the APs act as arbitrageurs who are more likely to transmit shocks from the ETF to the underlying assets. A policy that supports inventory-holding APs, or where government institutions themselves act as an AP, may help support the bond market when ETFs face large redemptions.

References

Andersen, Leif, Darrell Duffie, and Yang Song, "Funding Value Adjustments," *Journal of Finance*, 2019, *74*, 145–192.

Aramonte, Sirio and Fernando Avalos, "The recent distress in corporate bond markets: cues from ETFs," *BIS Bulletin No 6*, April 2020.

Bao, Jack, Jun Pan, and Jiang Wang, "The Illiquidity of Corporate Bonds," *Journal of Finance*, 2011, *66*, 911–946.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi, "Do ETFs Increase Volatilty?," *The Journal of Finance*, 2018, *73* (6), 2471–2535.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman, "A Survey of the Microstructure of Fixed-Income Markets," *Journal of Financial and Quantitative Analysis*, 2020, *55*.

_ , Kathleen M., William F. Maxwell, and Danielle Xu, "Measuring Abnormal Bond Performance," *Review of Financial Studies*, 2009, *22*, 4219–4258.

_ , Stacey Jacobsen, William Maxwell, and Kumar Venkataraman, "Capital Commitment and Illiquidity in Corporate Bonds," *Journal of Finance*, 2018, *73*, 1616–1661.

Bhattacharya, Ayan and Maureen O'Hara, "Can ETFs Increase Market Fragility? Effect of Information Linkages in ETF Markets," *Working Paper*, April 2018.

_ and _ , "ETFs and Systemic Risks," *CFA Institute Research Foundation Briefs*, January 2020.

BlackRock, "By the numbers: the bond ETF primary market process," 2021.

Boyarchenko, Nina, Anna Kovner, and Or Shachar, "It's What You Say and What You Buy: A Holistic Evaluation of the Corporate Credit Facilities," *FRB of New York Staff Report No. 935*, 2020.

Brogaard, Jonathan, Davidson Heath, and Da Huang, "The Rise of ETF Trading and the Bifurcation of Liquidity," *Working paper*, 2019.

Campbell, John Y., Sanford J. Grossman, and Jiang Wang, "Trading Volume and Serial Correlation in Stock Returns," *The Quartery Journal of Economics*, 1993, *108* (4), 905–939.

Cherkes, Martin, Jacob Sagi, and Richard Stanton, "A Liquidity-Based Theory of Closed-End Funds," *Review of Financial Studies*, 2009, *22* (1), 257–297.

Choi, Jaewon, Saeid Hoseinzadec, Sean Seunghun Shin, and Hassan Tehranian, "Corporate bond mutual funds and asset fire sales," *Journal of Financial Economics*, 2020, *138* (2), 432–457.

Coval, Joshua and Erik Stafford, "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Da, Zhi and Sophie Shive, "Exchange traded funds and asset return correlations. 24, 136–168," *European Financial Management*, 2018, *24*, 136–168.

D'Amico, Stefania, Vamsidhar Kurakula, and Stephen Lee, "Impacts of the Fed Corporate Credit Facilities through the Lenses of ETFs and CDX," *Working paper*, 2020.

Dannhauser, Caitlin, "The Impact of Innovation: Evidence from Corporate Bond ETFs," *Journal of Financial Economics*, 2017, *125*, 537–560.

Ellul, Andrew, Chotibhak Jotikasthira, and Christian Lundblad, "Regulatory pressure and fire sales in the corporate bond market," *Journal of Financial Economics*, 2011, *101* (3), 596–620.

Goldstein, Itay, Hao Jiang, and David T. Ng, "Investor flows and fragility in corporate bond funds," *Journal of Financial Economics*, 2017, *126* (3), 592–613.

Goldstein, Michael A. and Edith S. Hotchkiss, "Providing liquidity in an illiquid market: Dealer behavior in US corporate bonds," *Journal of Financial Economics*, 2020, *135* (1), 16–40.

Gorbatikov, Evgenii and Taisiya Sikorskaya, "Tow APs Are Better Than One: ETF Mispricing and Primary Market Participation," *Working Paper*, 2022.

Haddad, Valentin, Alan Moreira, and Tyler Muir, "When Selling Becomes Viral: Disruptions in Debt Markets in the Covid-19 Crisis and the Fed's Response," *Review of Financial Studies*, 2021, *34* (11), 5309–5351.

Holden, Craig and Jayoung Nam, "Market Accessibility, Corporate Bond ETFs, and Liquidity," *Working paper*, 2019.

Johnson, Steve, "Bond ETFs suck liquidity out of market in a crisis, academics say," 2023.

Koont, Naz, Yiming Ma, Lubos Pastor, and Yao Zeng, "Steering a Ship in Illiquid Waters: Active Management of Passive Funds," *Working paper*, 2022.

Lettau, Martin and Ananth Madhavan, "Exchange-Traded Funds 101 for Economists," *Journal of Economic Perspectives*, 2018, *32*, 135–154.

Ma, Yiming, Kairong Xiao, and Yao Zeng, "Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity," *The Review of Financial Studies*, 2022, *35* (10), 4674–4711.

Madhavan, Ananth and Aleksander Sobczyk, "Price Dynamics and Liquidity of Exchange-Traded Funds," *Journal of Investment Management*, 2016, *14* (2), 86–102.

Malamud, Semyon, "A dynamic equilibrium model of ETFs," *Working paper*, 2015.

O'Hara, Maureen and Xing (Alex) Zhou, "Anatomy of a liquidity crisis: Corporate bonds in the COVID-19 crisis," *Journal of Financial Economics*, 2021, *142* (1), 46–68.

Pagano, Marco, Antonio Sanchez Serrano, and Jozef Zechner, "Can ETFs contribute to systemic risk?," *Working Paper*, 2020.

Pan, Kevin and Yao Zeng, "ETF Arbitrage Under Liquidity Mismatch," *Working Paper*, 2021.

Pastor, Lubos and Robert F. Stambaugh, "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy*, 2003, *111* (3), 642–685.

Petajisto, Antti, "Inefficiencies in the Pricing of Exchange-Traded Funds," *Working Paper*, 2016.

Pramuk, Jacob, "Icahn blames BlackRock for bond market volatility," July 15 2015.

Ramaswamy, Srichander, "Market structures and systemic risks of exchange-traded funds," *BIS Working Paper No. 343*, 2011.

Reilly, Christopher, "The Hidden Cost of Corporate Bond ETFs," *Working paper*, 2021.

Saglam, Mehmet, Tugkan Tuzun, and Russ Wermers, "Do ETFs Increase Liquidity?," *Working paper*, 2019.

Shim, John J., "Arbitrage Comovement," *Working Paper*, 2020.

_ **and Karamfil Todorov**, "ETFs, illiquid assets, and fire sales," *BIS Working Paper No 975*, 2021.

Shleifer, Andrei and Robert Vishny, "Liquidation Values and Debt Capacity: A Market Equilibrium Approach," *Journal of Finance*, 1992, *47* (4), 1343–1366.

Sushko, Vladyslav and Grant Turner, "The implications of passive investing for securities markets," *BIS Quarterly Review, March*, March 2018.

Todorov, Karamfil, "Passive Funds Affect Prices: Evidence from the most ETF-dominated Markets," *Working Paper*, 2019.

_ , "The anatomy of bond ETF arbitrage," *BIS Quarterly Review*, March 2021.

Wigglesworth, Robin, "Jane Street: the top Wall Street firm "no one's heard of"," 2021.

Zuckerman, Gregory, Katy Burne, and Sarah Krouse, "https://www.wsj.com/articles/carl-icahn-fuels-criticism-of-bond-etfs-1437092991," July 16 2015.

ETFs, Illiquid Assets, and Fire Sales

[INTERNET APPENDIX]

John J. Shim and Karamfil Todorov

April 2023

A Proofs of Main Text Propositions

A.1 Proof of Proposition 1 (Redemption Equilibrium)

Proof. First, we solve for γ_c, which maximizes

$$P_e(\gamma) = P_0 - c(q+z) + cz\gamma - \left(\frac{\lambda}{2} - c\right) q\gamma^2 \tag{A.1}$$

assuming we hold price impact cost fixed at c (regardless of q). The derivative of $P_e(\gamma)$ with respect to γ is

$$\frac{dP_e}{d\gamma} = cz + 2\gamma q\left(c - \frac{\lambda}{2}\right).$$

If $\frac{\lambda}{2} < c$, $P_e(\gamma)$ is strictly increasing and convex from 0 to 1. Hence, the optimal solution is $\gamma_c = 1$. $P_e(\gamma)$ is concave if $c < \frac{\lambda}{2}$, and then there are three cases. First, when $z > q\frac{\lambda-2c}{c}$, $P_e(\gamma)$ is increasing from $\gamma = 0$ and $\gamma = 1$ and thus $\gamma_c = 1$. Second, when $z = 0$, $P_e(\gamma)$ is decreasing from $\gamma = 0$ to $\gamma = 1$ and thus $\gamma_c = 0$. Third, when $z > 0$ and $z < q\frac{\lambda-2c}{c}$, γ_c is between 0 and 1 . In this case, the first-order condition yields the solution, which is

$$\gamma_c = \frac{cz}{(\lambda - 2c)\, q}.$$

We now reintroduce the possibility of fire-sales. γ_c does not trigger a fire sale if $(1-\gamma_c)\,q \leq \tau$, and the γ_c found above is the optimal γ^*_{RD} since it maximizes the ETF price. γ_c triggers a fire sale if $(1-\gamma_c)\,q > \tau$. Thus, one must compare the maximum ETF price under fire sale conditions with the maximum ETF price when the AP sells up to the limit τ.

Let $P_e(\gamma; c)$ denote the ETF price from equation A.1 with price impact cost c, and $P_e(\gamma; f)$ the ETF price expression but with price impact cost of f instead of c. Denote γ_τ as the smallest holding fraction that does not trigger a fire sale, i.e., $\gamma_\tau = 1 - \frac{\tau}{q}$. In the case when $(1-\gamma_c)\,q > \tau$, $\gamma_\tau > \gamma_c$ and thus $P_e(\gamma_c; c) > P_e(\gamma_\tau; c)$ since a fire sale only occurs when P_e is decreasing from γ_c to 1. Thus, we need to analyze fire-sale scenarios when $0 \leq \gamma_c < 1$. The AP chooses γ_τ if it yields a greater ETF price than γ_f under a fire sale, $P_e(\gamma_\tau; c) > P_e(\gamma_f; f)$, where γ_f maximizes $P_e(\gamma; f)$ for the region $0 \leq \gamma \leq 1$.

There are two cases for $P_e(\gamma; f)$. The first is when $c < \frac{\lambda}{2} < f$, and then $P_e(\gamma; c)$ is concave, but $P_e(\gamma; f)$ is convex and increasing from 0 to 1. But since $P_e(1; c) = P_e(1; f)$ and $P_e(0; c) > P_e(0; f)$, $P_e(\gamma_\tau, c)$ is always greater than $P_e(\gamma; f)$ for any γ from 0 to 1. Thus, in this case $\gamma^*_{RD} = \gamma_\tau$.

The second case is when $c < f < \frac{\lambda}{2}$, and then both $P_e(\gamma; c)$ and $P_e(\gamma; f)$ are concave. If γ_c is interior and triggers a fire sale, then $0 < z < q\frac{\lambda - 2c}{c}$, $\frac{\lambda}{2} > c$, and $(1-\gamma_c)\,q > \tau$. Under these conditions, we must check that $P_e(\gamma_\tau; c) > P_e(\gamma; f)$ for any γ. $P_e(\gamma; f)$ is maximized at

$$\gamma_f = \frac{fz}{(\lambda - 2f)\,q}.$$

Thus, it is sufficient to show that $P_e(\gamma_f; f) < P_e(\gamma_\tau; c)$. We denote $\bar{\lambda}$ as the maximum possible λ that satisfies this condition. That is, as long as holding costs are at or below $\bar{\lambda}$, the AP avoids a fire sale and selects γ_τ instead of γ_f. $\bar{\lambda}$ is given by the positive root of the following quadratic equation:

$$\lambda^2 (q-\tau)^2 - \lambda\left(2f(q-\tau)^2 + Y\right) + 2fY + (fz)^2 = 0,$$

where

$$Y = 2q(f-c)(q+z) + 2(q-\tau)cz + 2c(q-\tau)^2,$$

which yields the condition

$$\bar{\lambda} = \frac{2f\left(q-\tau\right)^2 + Y + \sqrt{\left(2f\left(q-\tau\right)^2 + Y\right)^2 - 4\left(q-\tau\right)^2\left(2fY + \left(fz\right)^2\right)}}{2\left(q-\tau\right)^2}.$$

If there are cases where holdings costs λ are so large relative to fire sales costs f (i.e., $\lambda > \bar{\lambda}$) then the condition is not satisfied. These cases are perhaps less interesting from a practical point of view since fire sale costs are usually assumed to be large due to additional externalities (Shleifer and Vishny, 1992). If holding costs are above $\bar{\lambda}$, then the equilibrium γ^*_{RD} is given by γ_f, i.e., the AP sells at fire-sale prices. As long as $\lambda < \bar{\lambda}$, then in any fire-sale scenario where $(1-\gamma_c)\,q > \tau$ the AP maximizes the ETF price by selecting $\gamma^*_{RD} = \gamma_\tau$. In all other scenarios, γ_c does not trigger a fire sale, and thus $\gamma^*_{RD} = \gamma_c$.

To complete the proof, we provide expressions for the discount (negative premium) as a function of γ. After redeeming q ETF shares for q units of bond A, the equilibrium NAV of the ETF is

$$\begin{aligned} NAV^*_{RD} &= \frac{1}{1-q}\left(\left(0.5-q\right)\left(P_0 - \left(1-\gamma^*_{RD}\right)cq\right) + 0.5P_0\right) \\ &= P_0 - \frac{0.5-q}{1-q}\left(1-\gamma^*_{RD}\right)cq, \end{aligned}$$

the equilibrium ETF price is

$$P^*_{e,RD} = P_0 - c\left(q+z\right) + cz\gamma^*_{RD} - \left(\frac{\lambda}{2} - c\right)q\left(\gamma^*_{RD}\right)^2,$$

and the discount is

$$\begin{aligned} -\pi^*_{RD} &= -(P^*_{e,RD} - NAV^*_{RD}) \\ &= \left(\gamma^*_{RD}\right)^2\left(\frac{\lambda}{2} - c\right)q + \left(1-\gamma^*_{RD}\right)zc + \frac{0.5 + \gamma\left(0.5-q\right)}{1-q}cq. \end{aligned} \tag{A.2}$$

□

A.2 Proof of Proposition 2 (Creation Equilibrium)

Proof. We solve for γ^*_{CR}, which minimizes the ETF price for purchasing ETF investors. The AP's payoff is

$$(P_e - NAV_{basket}) \cdot q - \frac{\lambda}{2} \cdot (\gamma q)^2 + (z - \gamma q) \cdot \Delta P_A,$$

where bond A is included in the creation basket. Just as for redemptions, we assume the AP makes zero profits. Setting the AP's payoff to zero provides an expression for the ETF price

$$P_e = P_0 + \gamma c z - c\,(z - q) - \gamma^2 \left(c - \frac{\lambda}{2}\right) q.$$

There are two cases: $\lambda > 2c$ and $\lambda < 2c$.

If $\lambda > 2c$, $P_e(\gamma)$ is convex and increasing between 0 and 1. Thus, the minimum price is achieved at $\gamma^*_{CR} = 0$. This is easy to see by taking first-order conditions, which yields

$$\gamma = \frac{c}{(2c - \lambda)} \frac{z}{q}.$$

Since $\lambda > 2c$, $\gamma < 0$ and it is optimal to set $\gamma^*_{CR} = 0$.

If $\lambda < 2c$, $P_e(\gamma)$ is concave. Thus, we must compare $P_e\,(0)$ and the ETF price at the maximum possible γ. The maximum γ is given by $\bar{\gamma} \equiv \min\left(\frac{z}{q}, 1\right)$ since if $z < q$, the AP can only set γ up to $\frac{z}{q}$ (she holds z, thus can only take up to z from her inventory). For $\gamma = \frac{z}{q}$, we have that $P_e\,(0) < P_e\left(\frac{z}{q}\right)$ since $P_e\left(\frac{z}{q}\right) = P_e\,(0) + \frac{\lambda z^2}{2q} > P_e\,(0)$. Thus, in this case $\gamma^*_{CR} = 0$. If $\bar{\gamma} = 1$, we also have $P_e\,(0) < P_e\,(1)$. To see this, we expand $P_e\,(0) < P_e\,(1)$, which yields

$$q\left(1 - \frac{\lambda}{2c}\right) < z.$$

This must be satisfied. If the condition is not and $z \leq (1 - \frac{\lambda}{2c})q$, then it is also true that $z < q$ and it is impossible to set $\gamma = 1$ since $\bar{\gamma} = \min\left(\frac{z}{q}, 1\right)$, i.e., there is not enough inventory to take all q bonds from inventory. And, we have also just shown that $P_e\,(0) < P_e\left(\frac{z}{q}\right)$. Thus, $\gamma^*_{CR} = 0$.

In all cases, the AP chooses $\gamma^*_{CR} = 0$. The equilibrium NAV is given by

$$\begin{aligned} NAV^*_{CR} &= \frac{1}{1+q}\left((0.5+q)\left(P_0+cq\right)+0.5P_0\right) \\ &= P_0 + \frac{0.5+q}{1+q}cq, \end{aligned}$$

the equilibrium ETF price is

$$P^*_{e,CR} = P_0 - c\left(z-q\right),$$

and the equilibrium premium is

$$\begin{aligned} \pi^*_{CR} &= P^*_{e,CR} - NAV^*_{CR} \\ &= \frac{0.5}{1+q}cq - cz. \end{aligned} \tag{A.3}$$

□

A.3 Proof of Proposition 3 (Multiple APs with Heterogeneous Inventory)

Proof. We first start with the creation case. AP_1 and AP_2 each quote the lowest ETF price at which they are willing to sell. Proposition 2 shows that each AP's most competitive price is given by

$$P^*_{e,CR} = P_0 - c(z-q).$$

It is straightforward to see that the ETF price is the lowest for the AP that holds the most inventory, i.e., AP_2. Thus, AP_2 is always the marginal AP for creation, and the ETF price, NAV, and premium are given by Proposition 2 with inventory equal to z_2.

The marginal AP for redemption depends on z_1 and z_2. Since z_1 is strictly less than z_2, we start by showing who the marginal AP is for various values of z_1. Denote the ETF price of AP_1 and AP_2 as given by the function $P_{e,RD}\left(z\right)$, which depends on each AP's inventory level, and the function follows directly from Proposition 1. In addition, $\gamma_{RD,AP1}$ and $\gamma_{RD,AP2}$ correspond to the γ^*_{RD} given in Proposition 1 as if each AP were the marginal AP. For

redemption, the AP with the highest price ends up being the marginal AP.

There are six redemption cases.

(Case 1) If AP_1 and AP_2 each have sufficiently large inventory, such that $\frac{\lambda-2c}{c}q < z_1 < z_2$, then they quote identical prices and either can be marginal.

(Case 2) If AP_1 has no inventory ($z_1 = 0$ and $z_2 > 0$), then $\gamma_{RD,AP1} = 0$ and

$$P_{e,RD}(z_1) = P_0 - cq$$
$$P_{e,RD}(z_2) = P_0 - c(q + z_2) + \gamma_{RD,AP2} c z_2 - (\gamma_{RD,AP2})^2 \left(\frac{\lambda}{2} - c\right) q.$$

The difference between the two prices is

$$P_{e,RD}(z_1) - P_{e,RD}(z_2) = c z_2 (1 - \gamma_{RD,AP2}) + (\gamma_{RD,AP2})^2 \left(\frac{\lambda}{2} - c\right) q$$

and since $0 < \gamma_{RD,AP2} \leq 1$, $P_{e,RD}(z_1) - P_{e,RD}(z_2)$ is strictly positive (assuming $\lambda > 2c$ as in our baseline framework), i.e., AP_1 quotes the highest price and is the marginal AP.

(Case 3) If AP_1 has inventory such that $0 < z_1 < \frac{\lambda-2c}{c}q$ and AP_2 has sufficiently large inventory such that $\frac{\lambda-2c}{c}q < z_2$, then $0 < \gamma_{RD,AP1} < 1$ and $\gamma_{RD,AP2} = 1$. The price difference between the two APs is

$$P_{e,RD}(z_1) - P_{e,RD}(z_2) = c z_1 (\gamma_{RD,AP1} - 1) + \left(1 - (\gamma_{RD,AP1})^2\right) \left(\frac{\lambda}{2} - c\right) q.$$

Thus, AP_1 is marginal if

$$z_1 < \frac{1}{2} \frac{1 - (\gamma_{RD,AP1})^2}{1 - \gamma_{RD,AP1}} \frac{(\lambda - 2c)}{c} q,$$

which is satisfied if z_1 is sufficiently small relative to q. Otherwise, AP_2 is marginal.

(Case 4) If AP_1 and AP_2 both have inventory such that $0 < z_1 < z_2 < \frac{\lambda-2c}{c}q$, then both $\gamma_{RD,AP1}$ and $\gamma_{RD,AP2}$ are between 0 and 1. If both APs have "typical" (non-fire-sale) redemptions, then we use $\gamma_c(z) \equiv \gamma_c = \frac{cz}{(\lambda-2c)q}$, where γ_c is from Proposition 1, to plug in for

$\gamma_{RD,AP1}$ and $\gamma_{RD,AP2}$, which yields

$$\begin{aligned} P_{e,RD}(z_1) - P_{e,RD}(z_2) &= c(z_2 - z_1) + \frac{c^2(z_1^2 - z_2^2)}{2(\lambda - 2c)q} \\ &= c(z_2 - z_1)\left(1 - \frac{1}{2}\gamma_{RD,AP1} - \frac{1}{2}\gamma_{RD,AP2}\right), \end{aligned}$$

which is always positive since $0 < z_1 < z_2$ and $0 < \gamma_{RD,AP1} < \gamma_{RD,AP2} < 1$. Thus AP_1 is the marginal AP.

(Case 5) If AP_1 and AP_2 both have inventory such that $0 < z_1 < z_2 < \frac{\lambda - 2c}{c}q$ and both APs have fire-sale redemptions, then $\gamma_{RD,AP1} = \gamma_{RD,AP2} = \gamma_\tau$ and

$$P_{e,RD}(z_1) - P_{e,RD}(z_2) = (1 - \gamma_\tau)c(z_2 - z_1).$$

This expression is always positive since $0 < z_1 < z_2$ and $0 < \gamma_\tau < 1$, and AP_1 is marginal.

(Case 6) If AP_1 and AP_2 both have inventory such that $0 < z_1 < z_2 < \frac{\lambda - 2c}{c}q$, and AP_1 has a fire sale redemption and AP_2 has a typical redemption. The price difference is

$$P_{e,RD}(z_1) - P_{e,RD}(z_2) = -\frac{\tau}{q}cz_1 - \left(1 - \frac{\tau}{q}\right)^2\left(\frac{\lambda}{2} - c\right)q - \gamma_c(z_2)\frac{cz_2}{2} + cz_2.$$

AP_1 is marginal if

$$cz_2 - \frac{c^2 z_2^2}{2(\lambda - 2c)q} > \frac{\tau}{q}cz_1 + \left(1 - \frac{\tau}{q}\right)^2\left(\frac{\lambda}{2} - c\right)q.$$

Intuitively, this captures a nuanced case where AP_1 sells more aggressively (holds a smaller fraction of bonds) because she has less inventory, but this aggressive selling may cause her to run into the fire-sale threshold. As a result, she changes her behavior to minimize total costs of handling the redeemed bonds and holds more bonds because fire sales are costly, even with pure arbitrage incentives.

In four of the six possible redemption scenarios, AP_1 quotes the (weakly) largest ETF price. In Cases 3 and 6, the marginal AP depends on parameters. AP_1 is marginal in all the cases where AP_1 has no inventory. (Note: it is not possible for AP_1 to have a typical

redemption and AP_2 to have a fire sale redemption since this implies $(1 - \gamma_c(z_2))\, q > \tau$ and $(1 - \gamma_c(z_1))\, q \leq \tau$, but together this implies $z_1 > z_2$, which is a contradiction.) $\square$

B Full Basket Models

In this section, we analyze the case when baskets are identical to holdings ("full baskets") since this case helps us illustrate the effect of fractional baskets on discounts/premiums, and also allows us to compare the ETF with a stylized mutual fund (MF). The full baskets case is also useful to describe the case of Treasuries and equity ETFs, which empirically have close to full baskets (see Shim and Todorov (2021) and Todorov (2021)).

Note: the mutual fund case described in Section 3.6 is obtained by setting $\gamma = 0$ and $z = 0$ in the expressions below, which yields ETF prices equal to NAV ($P_{e,full} = NAV_{full}$), and premium/discount of zero ($\pi_{full} = 0$). This is true both for creations and redemptions.

B.1 Redemption with Full Baskets

To illustrate the effects on discounts parsimoniously, assume the ETF puts equal amounts of bonds A and B in the basket $q_A = q_B = \frac{q}{2}$, which corresponds to the proportion of the two bonds in holdings. The AP chooses $\gamma_A = \gamma_B = \gamma$, and the redemption is typical. We present the discount with full baskets below, and then compare it to Proposition 1 below. We omit RD subscripts below to simplify the notation.

Proposition 4. *(Redemption with Full Baskets) In the case of typical (non fire sale) redemptions with full baskets, the maximum ETF price is larger and the corresponding discount is smaller when compared to the case with fractional baskets. The ETF puts equal amounts of bonds A and B in the basket ($q_A = q_B = \frac{q}{2}$) and the AP chooses $\gamma_A = \gamma_B = \gamma$. The ETF price is*

$$P^*_{e,full} = P_0 - \frac{cq}{2} - zc + \gamma zc - \left(\frac{\lambda}{4} - \frac{c}{2}\right)\gamma^2 q,$$

the NAV of the ETF is

$$NAV^*_{full} = P_0 - \frac{c(1-\gamma)q}{2},$$

and the ETF discount is

$$-\pi^*_{full} = \left(\frac{\lambda}{4} - \frac{c}{2}\right)\gamma^2 q + \frac{c\gamma q}{2} + zc(1-\gamma).$$

Proof of Proposition 4.

Proof. Using $P_A = P_B = P_0$, we can write AP's payoff as:

$$\left(0.5\left(P_0 - c(1-\gamma)\frac{q}{2}\right) + 0.5\left(P_0 - c(1-\gamma)\frac{q}{2}\right) - P_e\right)q - \lambda\left(\frac{\gamma q}{2}\right)^2 - 2\left(z + \frac{\gamma q}{2}\right)c(1-\gamma)\frac{q}{2}.$$

Setting the AP's payoff to zero yields and expression for the ETF price

$$P_e = \left(\frac{c}{2} - \frac{\lambda}{4}\right)\gamma^2 q + \gamma z c + P_0 - \frac{cq}{2} - zc.$$

Taking FOC gives

$$\gamma_{full} = \frac{zc}{(\frac{\lambda}{2} - c)q} = 2\gamma_c.$$

The steps for this proof are similar to the proof for Proposition A.1 so we focus on the difference between γ^*_{RD} and the resulting price, NAV, and discount for full-basket counterparts. For this proof and for comparison purposes, we add the subscript "$full$" for the full basket objects and "$fractional$" for the fractional basket objects, and omit the subscript RD since both refer to the redemption scenario. For $0 < \gamma^*_{full} < 1$ where $\gamma^*_{full} = \gamma_{full}$, we need $0 < z < q\left(\frac{\lambda}{2c} - 1\right)$. For $z = 0$, the solution is $\gamma^*_{full} = 0$. Otherwise, the solution is $\gamma^*_{full} = 1$. The maximum ETF price is

$$P^*_{e,full} = P_0 - \frac{cq}{2} - zc + \gamma^*_{full} zc - \left(\frac{\lambda}{4} - \frac{c}{2}\right)\left(\gamma^*_{full}\right)^2 q, \tag{B.1}$$

the NAV is

$$\begin{aligned} NAV^*_{full} &= \frac{1}{1-q}\left(\left(0.5 - \frac{q}{2}\right)\left(P_0 - \frac{c\left(1-\gamma^*_{full}\right)q}{2}\right) + \left(0.5 - \frac{q}{2}\right)\left(P_0 - \frac{c\left(1-\gamma^*_{full}\right)q}{2}\right)\right) \\ &= P_0 - \frac{c\left(1-\gamma^*_{full}\right)q}{2}. \end{aligned}$$

The discount is

$$\begin{aligned}-\pi^*_{full} &= -(P_{e,full} - NAV_{full}) \\ &= \frac{cq\gamma^*_{full}}{2} + zc\left(1-\gamma^*_{full}\right) + \left(\frac{\lambda}{4} - \frac{c}{2}\right)\left(\gamma^*_{full}\right)^2 q.\end{aligned}$$

The optimal ETF price with full baskets is always larger than the optimal price with fractional baskets because $P_{e,full}\left(\gamma^*_{full}\right) > P^*_{e,full}\left(\gamma^*_{fractional}\right) > P^*_e\left(\gamma^*_{fractional}\right)$. The optimal discount with fractional baskets is also larger than the one with full baskets. Comparing the two discounts at the optimal γ's, when $0 \leq \gamma^*_{full} < 1$, gives:

$$\begin{aligned}-\pi^*_{fractional} - (-\pi^*_{full}) &= \gamma_c^2\left(\frac{\lambda}{2} - c\right)q + (1-\gamma_c)zc + \frac{0.5 + \gamma_c\left(0.5-q\right)}{1-q}cq \\ &\quad - \left[2\gamma_c^2\left(\frac{\lambda}{2} - c\right)q + \left(1-2\gamma_c\right)zc + c\gamma_c q\right] \\ &= -\gamma_c^2\left(\frac{\lambda}{2} - c\right)q + \gamma_c c\left(z - \frac{0.5q}{1-q}\right) + \frac{0.5}{1-q}cq \\ &= \frac{c^2z^2}{2\left(\lambda - 2c\right)q} + \frac{0.5}{1-q}cq\left(1 - \frac{cz}{\left(\lambda-2c\right)q}\right) > 0. \qquad \text{(B.2)}\end{aligned}$$

The expression above is positive because $cz < \left(\lambda - 2c\right)q$. When $\gamma^*_{full} = \gamma^*_{fractional} = 1$ (i.e., $\gamma_c \geq 1$), this implies that $\gamma_{full} \geq 1$. Thus,

$$\begin{aligned}-\pi^*_{fractional} - (-\pi^*_{full}) &= -\left(\frac{\lambda}{2} - c\right)q + c\left(z - \frac{0.5q}{1-q}\right) + \frac{0.5}{1-q}cq \\ &= cz - \left(\frac{\lambda - 2c}{2}\right)q > 0\end{aligned}$$

where the last line follows from $\gamma_c \geq 1$ (which yields $cz \geq \left(\lambda - 2c\right)q$). The last case is where $\gamma^*_{full} = 1$ and $0 < \gamma^*_{fractional} < 1$, which occurs when

$$\frac{\left(\lambda - 2c\right)}{2c}q < z < \frac{\left(\lambda - 2c\right)q}{c}. \qquad \text{(B.3)}$$

In this case, the difference in discounts is

$$-\pi^*_{fractional} - (-\pi^*_{full}) = \left[-\left(\frac{\lambda}{2} - c\right) q(2-\gamma_c^2) + zc(2-\gamma_c)\right] + (\gamma_c(0.5-q) + q - 0.5)\frac{cq}{1-q} > 0$$

since the first term is greater than zero from equation B.3 and the last term is greater than zero from the restriction $q < 0.5$. □

The intuition is that with full baskets, total holding costs are smaller (due to convexity, spreading the redeemed amount across two bonds decreases total costs), the price impact costs are smaller (since the arbitrage profit is concave in q), and the AP can quote a higher price, which decreases the discount.

In the context of our model, we interpret the discount as a type of "buffer" in the sense that a larger discount represents less selling pressure transmitted from the ETF to the underlying assets. Fractional baskets provide a larger buffer than full baskets for two reasons. First, holding costs are larger with fractional baskets since they are concentrated in one bond versus spread between two (and the cost is quadratic per bond). These larger costs are passed along to redeeming ETF investors in the form of a larger discount. Second, with fractional baskets, the AP concentrates selling in one bond instead of two, and that bond is precisely the one that has a lower weight in the ETF holdings after the redemption. That is, the ETF sheds bonds that are sold by the AP and overweights bonds that are not, which leads to a greater NAV.

B.2 Creation with Full Baskets

We now present the creation case with full baskets. To illustrate the effects of full baskets on premiums parsimoniously, assume again as in the case of redemptions that $q_A = q_B = q/2$ and the AP chooses $\gamma_A = \gamma_B = \gamma$. We omit CR subscripts below to simplify the notation.

Proposition 5. *(Creation with Full Baskets) In the case of creation with full baskets, the premium is smaller than with fractional baskets. The ETF puts equal amounts of bonds A*

and B in the basket ($q_A = q_B = \frac{q}{2}$) and the AP chooses $\gamma_A = \gamma_B = \gamma^*_{full} = 0$. The equilibrium ETF price is

$$P^*_{e,full} = P_0 - c\left(z - \frac{q}{2}\right),$$

the equilibrium NAV of the ETF is

$$NAV^*_{full} = P_0 + \frac{cq}{2},$$

and the equilibrium ETF premium is

$$\pi^*_{full} = -cz.$$

Proof of Proposition 5.

Proof. Using $P_A = P_B = P_0$, we can write the AP's payoff from creation with full baskets as

$$\left(P_e - P_0 - c \cdot (1-\gamma)\frac{q}{2}\right) \cdot q - \lambda \cdot \left(\frac{\gamma q}{2}\right)^2 + \frac{1}{2}(2z - \gamma q) \cdot c \cdot (1-\gamma)\, q.$$

Setting this equal to zero yields the expression for ETF price, or

$$P_e = P_0 - c\left(z - \frac{q}{2}\right) - \gamma^2\left(c - \frac{\lambda}{2}\right)\frac{q}{2} + \frac{1}{2}c\gamma\,(q + 2z)\,.$$

Again, the steps for this proof are similar to the proof for Proposition A.2 so we succinctly run through the steps. For this proof and for comparison purposes, we add the subscript "$full$" for the full basket objects and "$fractional$" for the fractional basket objects, and omit "CR" to simplify notation. The solution to the minimization problem is interior ($0 < \gamma^* < 1$) if $\lambda > 2c$, since then the function $P_e(\gamma)$ is convex. The optimal γ is then found by taking FOC:

$$\gamma_{full} = \frac{c}{2c - \lambda}\frac{2z + q}{q} \tag{B.4}$$

and $\gamma_{full} < 0$ since $\lambda > 2c$. Since $\gamma_{full} < 0$, the solution for this case is $\gamma^*_{full} = 0$ because the function $P_e(\gamma)$ is increasing from 0 to 1. Thus, the solution in this case is the same for creations with full and with fractional baskets.

If $\lambda < 2c$, the function $P_e(\gamma)$ is concave, and the solution is again $\gamma = 0$ because $P_{e,full}(\gamma = 0) < P_{e,full}(\gamma = 1)$. If $z < q$, the price at $\gamma = \frac{z}{q}$ (using all the available inventory to create) is still larger than the price at $\gamma = 0$ since

$$P_{e,full}(0) = -c\left(z - \frac{q}{2}\right) + P_0,$$

whereas

$$P_{e,full}\left(\frac{z}{q}\right) = P_e(0) + \frac{\lambda z^2}{4q} + \frac{cz}{2} + \frac{cz^2}{2q} > P_e(0).$$

Thus, in this case $\gamma^*_{full} = 0$. The equilibrium ETF price is then

$$P^*_{e,full} = P_0 - c\left(z - \frac{q}{2}\right),$$

the equilibrium NAV is

$$\begin{aligned} NAV^*_{full} &= \frac{1}{1+q}\left(\left(0.5 + \frac{q}{2}\right)\left(P_0 + \frac{cq}{2}\right) + \left(0.5 + \frac{q}{2}\right)\left(P_0 + \frac{cq}{2}\right)\right) \\ &= P_0 + \frac{cq}{2}, \end{aligned}$$

and the equilibrium premium is

$$\pi^*_{full} = -cz.$$

□

The proposition shows that just like with fractional baskets, the creation equilibrium γ^*_{full} is also 0 with full baskets. The premium with full baskets is then negative, i.e., creations happen at a discount. The premium is smaller for full baskets compared to fractional baskets since $\frac{0.5}{1+q}cq - cz > -cz$. When the AP has no inventory, the premium is 0 with full baskets but positive with fractional baskets.

C Derivative of Premium/Discount with Respect to z

In this section, we compute the derivative of the discount and premium with respect to z.

Redemptions

For redemptions, there are two cases: the typical case where $(1-\gamma^*_{RD})\,q \leq \tau$, and the fire sale case.

- Fire Sale Case where $(1-\gamma_c)\,q > \tau$ and $\gamma^*_{RD} = \gamma_\tau$. It is trivial to take the derivative of the discount in equation A.2 when $\gamma^*_{RD} = \gamma_\tau$, which yields

$$\frac{-d\pi}{dz} = \frac{\tau c}{q} > 0.$$

- Typical Case where $(1-\gamma_c)\,q \leq \tau$ and $\gamma^*_{RD} = \gamma_c$. There are three sub-cases: when the AP holds all redeemed bonds ($\gamma_c = 1$), when the AP sells all redeemed bonds ($\gamma_c = 0$), and when the AP holds some fraction ($0 < \gamma_c < 1$). We analyze each case below.

  - AP holds all redeemed bonds ($\gamma_c = 1$). It is trivial to take the derivative of the discount from equation A.2 when $\gamma^*_{RD} = 1$, which yields

$$\frac{-d\pi}{dz} = 0.$$

  - AP sells all redeemed bonds ($\gamma_c = 0$). It is again trivial to take the derivative of equation A.2 when $\gamma^*_{RD} = 0$, which yields

$$\frac{-d\pi}{dz} = c > 0.$$

  - AP holds fraction of redeemed bonds ($0 < \gamma_c < 1$). Note that in this case $z < q\frac{(\lambda-2c)}{c}$. The discount in equation A.2 when $\gamma^*_{RD} = \gamma_c$ can be expressed as

$$-\pi_{RD} = -\frac{(cz)^2}{2\,(\lambda-2c)\,q} + cz + \frac{1}{2}\frac{cq}{1-q} + \frac{c^2 z(0.5-q)}{(\lambda-2c)\,(1-q)}.$$

Taking the derivative with respect to z yields

$$\frac{-d\pi}{dz} = \frac{c\left[c\left(\frac{1}{2}-q\right)q - c\left(1-q\right)z + \left(1-q\right)\left(\lambda-2c\right)q\right]}{\left(1-q\right)\left(\lambda-2c\right)q}.$$

Since $z < q\frac{(\lambda-2c)}{c}$ and $q < 0.5$, we have that $c\left(\frac{1}{2}-q\right)q - c\left(1-q\right)z + \left(1-q\right)\left(\lambda-2c\right)q > 0$, and thus $\frac{-d\pi}{dz} > 0$.

Creations

For creations, when the AP does not hold inventory in the creation basket bonds ($z = 0$), it is straightforward to show that $\frac{d\pi}{dz} = 0$ from equation A.3. When the AP holds inventory, she still chooses to set $\gamma^*_{CR} = 0$ and from equation A.3, it is trivial to show that $\frac{d\pi}{dz} = -c$.

D Creation Model Extensions

D.1 Creation with Mark-to-Market Multiplier

To account for the asymmetric effects of propping up mark-to-market values of inventory, we multiply the mark-to-market term in AP's payoff by a multiplier m, which accounts for the costs and benefits of adjusting mark-to-market values since $0 \leq m \leq 1$. The case of $m = 1$ captures the situation when mark-to-market gains are as beneficial for creations as for redemptions, whereas the case of $m = 0$ captures the situation when these motives are absent from AP's optimization problem as in the standard frictionless ETF arbitrage case. For simplicity, we only focus on the case $\lambda > 2mc$. Again, we omit the subscript "CR" to simplify the notation.

Proposition 6. *(Creation with Mark-to-Market Multiplier) In the case of creation with fractional baskets and mark-to-market multiplier m (where $\lambda > 2mc$), the equilibrium share of created bonds that is taken from AP's inventory, γ_m^* is given by:*

Proposition.

$$\gamma_m^* = \begin{cases} 0, & \text{if } z \geq \frac{q(1-m)}{m}, \text{ or } z = 0 \text{ , or } \frac{2cq(1-m)}{\lambda} \leq z < q \\ \frac{qc(1-m)-czm}{q(\lambda-2mc)}, & \text{if } \frac{q(1-m)}{m} - q\left(\frac{\lambda}{mc} - 2\right) \leq z < \frac{q(1-m)}{m} \\ 1, & \text{if } q \leq z < \frac{q(1-m)}{m} - q\left(\frac{\lambda}{mc} - 2\right) \\ \frac{z}{q}, & \text{if } 0 < z < \frac{2cq(1-m)}{\lambda} < q. \end{cases}$$

The ETF price is

$$P_{e,m}^* = P_0 - c\left(zm - q\right) - \left(\gamma_m^*\right)^2 \left(mc - \frac{\lambda}{2}\right) q + \gamma_m^* cmz - \gamma_m^* qc\left(1 - m\right),$$

the NAV of the ETF is

$$NAV_m^* = P_0 + \frac{0.5 + q}{1 + q}\left(1 - \gamma_m^*\right) cq, \tag{D.1}$$

and the ETF premium is

$$\pi_m^* = -c(zm-q) - (\gamma_m^*)^2\left(mc - \frac{\lambda}{2}\right)q + \gamma_m^* cmz - \gamma_m^* qc\,(1-m) - \frac{0.5+q}{1+q}\,(1-\gamma_m^*)\,cq.$$

Proof of Proposition 6.

Proof. AP's payoff is

$$\underbrace{(P_e - NAV_{basket})\cdot q}_{\text{arbitrage profit}} - \underbrace{\frac{\lambda}{2}\cdot(\gamma q)^2}_{\text{balance sheet cost}} + \underbrace{m(z-\gamma q)\cdot\Delta P_A}_{\text{mark-to-market gain}}$$

Setting the payoff equal to zero, we obtain the ETF price

$$P_e = \begin{cases} P_0 + \frac{\lambda}{2}\cdot q & ,\gamma = 1 \\ P_0 - c(zm-q) & ,\gamma = 0 \\ P_0 - c(zm-q) - \gamma^2\left(mc - \frac{\lambda}{2}\right)q + \gamma cmz - \gamma qc(1-m) & ,0<\gamma<1 \end{cases} \tag{D.2}$$

The solution to the minimization problem is interior ($0 < \gamma^* < 1$) because for $\lambda > 2mc$, the function $P_e(\gamma)$ is convex. The optimal γ is then found by taking FOC, which yields

$$\gamma_m = \frac{qc(1-m) - czm}{q(\lambda - 2mc)} \tag{D.3}$$

The optimal γ_m^* depends then on whether the parameters of the model give $0 < \gamma_m < 1$. AP's inventory needs to satisfy then:

$$\frac{q(1-m)}{m} - q\left(\frac{\lambda}{mc} - 2\right) < z < \frac{q(1-m)}{m}.$$

If $z \geq \frac{q(1-m)}{m}$, $\gamma_m^* < 0$, and the optimal solution is $\gamma_m^* = 0$ since the function $P_e(\gamma)$ is increasing from 0 to 1 and thus $P_e(\gamma = 0) < P_e(\gamma = 1)$. If $z < \frac{q(1-m)}{m} - q(\frac{\lambda}{mc} - 2)$, the optimal solution is $\gamma_m^* = 1$ since the function $P_e(\gamma)$ is decreasing from 0 to 1. Obviously, this

is a feasible solution if $z \geq q$. If $z < q < \frac{q(1-m)}{m} - q(\frac{\lambda}{mc} - 2)$, the optimal solution is to use as much inventory as possible and to set $\gamma_m^* = \frac{z}{q}$ if $z < \frac{2cq(1-m)}{\lambda}$ since then $P_e(\frac{z}{q}) < P_e(0)$. More formally,

$$\begin{aligned} P_e\left(\frac{z}{q}\right) &< P_e(0) \\ -\frac{z^2}{q^2}(mc - \frac{\lambda}{2})q + \frac{z^2}{q}cm - zc(1-m) &< 0 \\ z &< \frac{2cq(1-m)}{\lambda}. \end{aligned}$$

If $z \geq \frac{2cq(1-m)}{\lambda}$, then the optimal solution is to set $\gamma = 0$. Then, the optimal γ_m^* is given by:

$$\gamma_m^* = \begin{cases} 0, & \text{if } z \geq \frac{q(1-m)}{m}, \text{ or } z = 0 \text{ , or } \frac{2cq(1-m)}{\lambda} \leq z < q \\ \frac{qc(1-m)-czm}{q(\lambda-2mc)}, & \text{if } \frac{q(1-m)}{m} - q(\frac{\lambda}{mc} - 2) \leq z < \frac{q(1-m)}{m} \\ 1, & \text{if } q \leq z < \frac{q(1-m)}{m} - q(\frac{\lambda}{mc} - 2) \\ \frac{z}{q}, & \text{if } 0 < z < \frac{2cq(1-m)}{\lambda}. \end{cases}$$

The equilibrium NAV is (using bond A in the basket)

$$NAV_m^* = \begin{cases} \frac{1}{1+q}(0.5P_0 + (0.5+q)P_0) = P_0, & \gamma_m^* = 1 \\ P_0 + \frac{0.5+q}{1+q}cq, & \gamma_m^* = 0 \\ P_0 + \frac{0.5+q}{1+q}\left(1 - \gamma_m^*\right)cq, & 0 < \gamma_m^* < 1. \end{cases}$$

The premium is

$$\pi_m^* = \begin{cases} \frac{\lambda}{2} \cdot q, & \gamma_m^* = 1 \\ \frac{0.5}{1+q}cq - cmz, & \gamma_m^* = 0 \\ -c(zm - q) - \left(\gamma_m^*\right)^2\left(mc - \frac{\lambda}{2}\right)q + \gamma_m^* cmz & \\ \quad -\gamma_m^* qc(1-m) - \frac{0.5+q}{1+q}(1-\gamma_m^*)cq, & 0 < \gamma_m^* < 1. \end{cases}$$

□

Proposition 6 shows that the premium can be positive or negative. If m is small, the premium is more likely to be positive for $\gamma = 0$. However, if mark-to-market incentives are important also for creations and inventory is large, the premium can be negative: i.e., creations happen at a discount.

An interesting case is when there are no mark-to-market gains from a creation: $m = 0$. In that case, the unique solution is $\gamma_m^* = \frac{c}{\lambda}$, since the AP trades off price impact costs of pushing bond prices and making lower arbitrage profit vs. holding costs of deviating from the target inventory (this is easy to see by plugging $m = 0$ and taking the FOC in the general equation for the price in equation D.2). Then, π is positive.

D.2 Creation with Balance Sheet Gains and Multiplier

Taking bonds from the balance sheet to create ETF shares could be beneficial for the AP. The intuition is that putting bonds on balance sheet is costly, whereas disposing of bonds might result in gains since it frees up balance sheet space. In that case, the effect of holding costs goes in the opposite direction for creations and redemptions. In addition, to account for the asymmetric effects of holding gains and mark-to-market gains for creations vs redemptions, we also multiply those terms with $0 \leq m \leq 1$. We omit the subscript "CR" to simplify the notation.

Proposition 7. *(Creation with multiplier for holding gains and mark-to-market gains) In case of creation with fractional baskets and holding gains with a multiplier m, the equilibrium share of created bonds that is taken from AP's inventory, $\gamma^*_{m,gains}$ is given by:*

$$\gamma^*_{m,gains} = \begin{cases} 0, & \text{if } z > (\frac{\lambda}{2c} + \frac{1}{m})q \\ 1 & \text{if } q \leq z < (\frac{\lambda}{2c} + \frac{1}{m})q \\ \frac{z}{q}, & \text{if } z < q \end{cases}$$

The ETF price is

$$P^*_{e,m,gains} = P_0 - c(mz - q) - m(\gamma^*_{m,gains})^2(c + \frac{\lambda}{2})q + \gamma^*_{m,gains}c(mz - q(1-m)),$$

the NAV of the ETF is

$$NAV^*_{m,gains} = P_0 + \frac{0.5+q}{1+q}(1-\gamma^*_{m,gains})cq,$$

and the ETF premium is

$$\begin{aligned}\pi^*_{m,gains} &= P^*_{e,m,gains} - NAV^* \\ &= -c(mz-q) - m\left(\gamma^*_{m,gains}\right)^2\left(c+\frac{\lambda}{2}\right)q \\ &\quad + \gamma^*_{m,gains}c(mz-q(1-m)) - \frac{0.5+q}{1+q}(1-\gamma^*_{m,gains})cq.\end{aligned}$$

Proof of Proposition 7.

Proof. The AP's payoff is

$$\underbrace{(P_e - NAV_{basket})\cdot q}_{\text{arbitrage profit}} + m\left[\underbrace{\frac{\lambda}{2}\cdot(\gamma q)^2}_{\text{balance sheet gain}} + \underbrace{(z-\gamma q)\cdot\Delta P_A}_{\text{mark-to-market gain}}\right].$$

Setting the payoff to zero yields the ETF price

$$P_e = -m\gamma^2(c+\frac{\lambda}{2})q + \gamma mcz - c(zm-q) - \gamma cq(1-m) + P_0. \tag{D.4}$$

The function is concave with a maximum achieved at $\gamma_{m,gains} = \frac{c(mz-q(1-m))}{mq(2c+\lambda)}$. The solution to the minimization problem is then a corner one at $\gamma^*_{m,gains} = 0$ as long as

$$P_e(\gamma=0) < P_e(\gamma=1) \iff z > (\frac{\lambda}{2c}+\frac{1}{m})q.$$

If $z < (\frac{\lambda}{2c}+\frac{1}{m})q$, then the optimal solution is to use as much inventory as possible to create: $\gamma^*_{m,gains} = \frac{z}{q}$, if $z < q$ and $\gamma^*_{m,gains} = 1$ if $q \le z < (\frac{\lambda}{2c}+\frac{1}{m})q$. Note that $P_e(\gamma=\frac{z}{q}) < P_e(\gamma=0)$ because

$$P_e(\gamma=\frac{z}{q}) < P_e(\gamma=0) \iff -c(mz-q) - \frac{z^2}{q}(m\frac{\lambda}{2} + c(1-m)) < -c(mz-q),$$

which is always true. Then, the optimal $\gamma^*_{m,gains}$ is given by

$$\gamma^*_{m,gains} = \begin{cases} 0, & \text{if } z > \left(\frac{\lambda}{2c} + \frac{1}{m}\right) q \\ 1 & \text{if } q \leq z < \left(\frac{\lambda}{2c} + \frac{1}{m}\right) q \\ \frac{z}{q}, & \text{if } z < q \end{cases}$$

The ETF price is

$$P^*_{e,m,gains} = \begin{cases} P_0 - \frac{m\lambda}{2} q & , \gamma = 1 \\ P_0 - c(mz - q) & , \gamma = 0 \\ P_0 - c(mz - q) - m\gamma^2 \left(c + \frac{\lambda}{2}\right) q + \gamma c(mz - q(1-m)) & , 0 < \gamma < 1 \end{cases} \tag{D.5}$$

NAV is

$$NAV^*_{m,gains} = \begin{cases} \frac{1}{1+q}(0.5P_0 + (0.5+q)P_0)) = P_0 & , \gamma = 1 \\ P_0 + \frac{0.5+q}{1+q} cq & , \gamma = 0 \\ P_0 + \frac{0.5+q}{1+q}(1-\gamma)cq & , 0 < \gamma < 1 \end{cases}$$

The premium is

$$\pi^*_{m,gains} = \begin{cases} -\frac{m\lambda}{2} q & , \gamma = 1 \\ \frac{0.5}{1+q} cq - mcz & , \gamma = 0 \\ -c(mz - q) - m\gamma^2 \left(c + \frac{\lambda}{2}\right) q + \gamma c(mz - q(1-m)) - \frac{0.5+q}{1+q}(1-\gamma)cq & , 0 < \gamma < 1 \end{cases}$$

□

The intuition for setting $\gamma^*_{m,gains} = 0$ is that with a large inventory ($z > (\frac{\lambda}{2c} + \frac{1}{m})q$), the AP does not take any bonds from her balance sheet and instead benefits from pumping the mark-to-market value of her inventory by buying all bonds on the market. The intuition for $\gamma^*_{m,gains} = 1$ and $\gamma^*_{m,gains} = \frac{z}{q}$ is that with a smaller inventory, the benefits from freeing up balance sheet space and not propping up bond prices (which decreases arbitrage profits) outweigh the benefit from pushing up the mark-to-market value of the existing inventory.

Thus, the AP uses the whole inventory for creations.

The premium is negative if $m = 1$ (i.e., creation happens at a discount) when $\gamma^*_{m,gains} = 0$ (since $z > (\frac{\lambda}{2c} + 1)q > q$) and $\gamma^*_{m,gains} = 1$ since the AP is ready to accept negative arbitrage profits reflected in the negative premium because the holding gains or the mark-to-market gains compensate for the arbitrage loss. If balance sheet and mark-to-market considerations play less important role for creations than for redemptions and $m \sim 0$, the premium can also be positive.

E Empirics Appendix

E.1 Trading Volume in Creation vs. Redemption: Additional Regressions

In this appendix, we estimate regressions similar to the ones from Section 4.2 but with standardized volume:

$$V^{sd}_{i,e,t} = \alpha_{e,t} + \alpha_i + \Delta_{CR} \cdot \text{in\_cr\_basket}_{i,e,t} + \Delta_{RD} \cdot \text{in\_rd\_basket}_{i,e,t} + \varepsilon_{i,e,t},$$

where $V^{sd}_{i,e,t}$ is the standardized volume of bond i held by ETF e on a day t with creation or redemption flows. For creation, we combine bond volume from the day of creation and the day *before*; for redemption, we combine volume from the day of redemption and the day *after*. We do so to account for slight differences in the timing of AP trading and ETF flows. For example, a creation basket may be acquired by buying bonds the day before and the day of ETF creation, whereas redemption bonds could be sold on the day of redemption and the day after it.[41] The variable $\text{in\_cr\_basket}_{i,e,t}$ ($\text{in\_rd\_basket}_{i,e,t}$) equals 1 when bond i held by ETF e is included in the creation (redemption) basket on date t. Since the estimation involves all bond ETFs in our sample (and thus potentially different types of bonds), we also standardize each bond's volume by its standard deviation. This adjustment allows us to compare variation in volume across bonds since a bond with greater average volume also typically has greater variation in volume.

As mentioned in the main text, we include ETF×date and bond fixed effects. This allows us to isolate variation within an ETF on a particular day, and control for time-invariant bond characteristics. We also control for each bond's bid-ask spread, the bond's weight in the ETF, and the bond's trading volume over the past week (excluding the most recent day). We cluster standard errors at the bond level and the ETF×date level.

The model predicts that creation basket bonds should have greater volume compared to non-basket bonds since APs always buy the basket bonds in the market. This means

[41] We find qualitatively similar results using only volume on the day of the flow, as well as when we combine volume on the day before, the day of, and the day after creation/redemption.

Table 5: Basket Trading Volume

Dependent Variable:	(1) $V^{sd}_{i,e,t}$	(2) $V^{sd}_{i,e,t}$
in_cr_basket$_{i,e,t}$	0.108***	0.094***
	(34.066)	(24.665)
in_rd_basket$_{i,e,t}$	0.058***	0.050***
	(13.556)	(12.142)
ETF×Date F.E.	✓	✓
Bond F.E.	✓	✓
Controls		✓
R^2	0.076	0.093
Observations	9,189,742	9,189,742

Notes. This table presents estimates from regressing trading volume for bonds held by ETFs on flow days, on creation/redemption variables. For specification (1), we regress standardized bond volume on an indicator variable for whether the bond is in the creation or redemption basket. We include high-dimensional ETF×date fixed effects, which allow us to compare trading volume for bonds included in a basket against other bonds held by the ETF on the same day but not included in the basket. The dependent variable includes trading volume on the day of the flow and the day before (after) for creation (redemption) to allow for slight differences in the timing of ETF flows and trading in the underlying bonds. In specification (2), we include additional controls: the bond's bid-ask spread, weight in the ETF, and trading volume over the previous 5 days before creation/redemption. The sample runs from May 2017 to February 2020, and standard errors are clustered at the bond and ETF×date level, and t-stats are shown in parentheses. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section E.1 for more details.

that the estimate of Δ_{CR} should be positive. The model also predicts that the difference in volume between basket and non-basket bonds for redemption should be smaller than for creation. This means that the estimate of Δ_{RD} should be smaller than Δ_{CR}.

Table 5 provides the regression estimates for the two regressions described above, and supports the predictions of the model. First, in specifications (1) and (2), the estimates for Δ_{CR} (coefficient on in_cr_basket$_{i,e,t}$) show that creation basket bond volume is about 0.11 standard deviations greater than non-basket bond volume. This estimate is about twice as large as the one for redemptions, where Δ_{RD} (coefficient on in_rd_basket$_{i,e,t}$) is around 0.06 standard deviations, consistent with the model's prediction that $\Delta_{CR} > \Delta_{RD}$.

E.2 Premiums/Discounts as Trading Costs: Additional Regressions

In this appendix, we present robustness tests from Section 4.5. Specifically, we examine cumulative ETF and NAV returns over future trading days following a creation or redemp-

tion. We separately analyze a sample of creation days with a premium and redemption days with a discount. For each sample, we regress the h-day ETF and NAV returns following a creation or redemption on the ETF premium/discount, while controlling for the ETF return and NAV return on the creation/redemption day, or

$$r^{etf}_{e,t\to t+h} = \alpha_e + \alpha_t + \beta_{etf}\pi_{e,t} + \gamma_{etf} r^{etf}_{e,t} + \eta_{etf} r^{nav}_{e,t} + \varepsilon^{etf}_{e,t},$$
$$r^{nav}_{e,t\to t+h} = \alpha_e + \alpha_t + \beta_{nav}\pi_{e,t} + \gamma_{nav} r^{etf}_{e,t} + \eta_{nav} r^{nav}_{e,t} + \varepsilon^{nav}_{e,t},$$

where $r^{etf}_{e,t\to t+h}$ and $r^{nav}_{e,t\to t+h}$ are the cumulative ETF return and NAV return for ETF e over h days following the creation/redemption day t, $\pi_{e,t}$ is the ETF premium, $r^{etf}_{e,t}$ and $r^{nav}_{e,t}$ are the ETF and NAV return on day t, and α_e and α_t are ETF and day fixed effects. We cluster standard errors at the date and ETF level. We estimate the regression using horizons of $h = 1$ day (Table 6), $h = 10$ days (Table 7), and $h = 20$ days (Table 8).

Table 6: Premiums/Discounts and Returns, 1-Day Horizon

(a) Premiums

	(1)	(2)	(3)	(4)
Dependent Variable:	$r_{e,t\to t+1}$	$r_{e,t\to t+1}$	$r^{nav}_{e,t\to t+1}$	$r^{nav}_{e,t\to t+1}$
$\pi_{e,t}$	-0.739***	-0.557***	0.032	0.012
	(-10.599)	(-9.567)	(1.235)	(0.389)
$r_{e,t}$		-0.093*		0.095***
		(-1.858)		(3.028)
$r^{nav}_{e,t}$		0.247***		-0.018
		(4.273)		(-0.521)
R^2	0.680	0.690	0.467	0.472
Observations	6,113	6,113	6,113	6,113

(b) Discounts

	(5)	(6)	(7)	(8)
Dependent Variable:	$r_{e,t\to t+1}$	$r_{e,t\to t+1}$	$r^{nav}_{e,t\to t+1}$	$r^{nav}_{e,t\to t+1}$
$\pi_{e,t}$	-0.393***	-0.414***	0.249***	0.150*
	(-4.302)	(-3.832)	(3.335)	(1.920)
$r_{e,t}$		0.034		0.162*
		(0.341)		(1.973)
$r^{nav}_{e,t}$		-0.017		-0.100
		(-0.123)		(-1.400)
R^2	0.686	0.686	0.665	0.670
Observations	957	957	957	957

Notes. This table presents estimates from regressing cumulative ETF returns $r_{e,t\to t+1}$ and NAV returns $r^{nav}_{e,t\to t+1}$ over the day following an ETF flow, on the ETF premium. We present the estimates separately for creation days with a positive premium (Panel (a)) and redemption days with a discount (Panel (b)). In some specifications, we include the ETF return and NAV return on the day of the flow as controls. All regressions also include date and ETF fixed effects. The sample runs from May 2017 to February 2020, standard errors are clustered at the ETF and date levels, and t-stats are shown in parentheses. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section 4.5 for more details.

Table 7: Premiums/Discounts and Returns, 10-Day Horizon

(a) Premiums

	(1)	(2)	(3)	(4)
Dependent Variable:	$r_{e,t\to t+10}$	$r_{e,t\to t+10}$	$r^{nav}_{e,t\to t+10}$	$r^{nav}_{e,t\to t+10}$
$\pi_{e,t}$	-0.770***	-0.690***	0.034	0.006
	(-19.870)	(-6.426)	(1.137)	(0.055)
$r_{e,t}$		0.031		0.158
		(0.281)		(1.201)
$r^{nav}_{e,t}$		0.116		-0.023
		(0.877)		(-0.154)
R^2	0.533	0.534	0.473	0.474
Observations	6,107	6,107	6,107	6,107

(b) Discounts

	(5)	(6)	(7)	(8)
Dependent Variable:	$r_{e,t\to t+10}$	$r_{e,t\to t+10}$	$r^{nav}_{e,t\to t+10}$	$r^{nav}_{e,t\to t+10}$
$\pi_{e,t}$	-0.727	-1.235**	0.042	-0.515
	(-1.565)	(-2.537)	(0.084)	(-1.048)
$r_{e,t}$		0.845**		0.922**
		(2.465)		(2.472)
$r^{nav}_{e,t}$		-0.625**		-0.600**
		(-0.251)		(-2.068)
R^2	0.833	0.837	0.792	0.798
Observations	954	954	954	954

Notes. This table presents estimates from regressing cumulative ETF returns $r_{e,t\to t+10}$ and NAV returns $r^{nav}_{e,t\to t+10}$ over the 10 days following an ETF flow, on the ETF premium. We present the estimates separately for creation days with a positive premium (Panel (a)) and redemption days with a discount (Panel (b)). In some specifications, we include the ETF return and NAV return on the day of the flow as controls. All regressions also include date and ETF fixed effects. The sample runs from May 2017 to February 2020, standard errors are clustered at the ETF and date levels, and t-stats are shown in parentheses. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section 4.5 for more details.

Table 8: Premiums/Discounts and Returns, 20-Day Horizon

(a) Premiums

	(1)	(2)	(3)	(4)
Dependent Variable:	$r_{e,t\to t+20}$	$r_{e,t\to t+20}$	$r^{nav}_{e,t\to t+20}$	$r^{nav}_{e,t\to t+20}$
$\pi_{e,t}$	-0.814***	-0.741***	-0.004	-0.119
	(-13.939)	(-3.067)	(-0.057)	(-0.505)
$r_{e,t}$		-0.005		0.191
		(-0.025)		(0.866)
$r^{nav}_{e,t}$		0.102		-0.142
		(0.364)		(-0.516)
R^2	0.763	0.763	0.692	0.692
Observations	6,102	6,102	6,102	6,102

(b) Discounts

	(5)	(6)	(7)	(8)
Dependent Variable:	$r_{e,t\to t+20}$	$r_{e,t\to t+20}$	$r^{nav}_{e,t\to t+20}$	$r^{nav}_{e,t\to t+20}$
$\pi_{e,t}$	-1.187*	-1.354**	0.045	-0.242
	(-1.906)	(-2.380)	(0.060)	(-0.380)
$r_{e,t}$		0.288		0.487
		(0.516)		(0.945)
$r^{nav}_{e,t}$		-0.348		-0.507
		(-0.894)		(-1.394)
R^2	0.893	0.893	0.900	0.901
Observations	952	952	952	952

Notes. This table presents estimates from regressing cumulative ETF returns $r_{e,t\to t+20}$ and NAV returns $r^{nav}_{e,t\to t+20}$ over the 20 days following an ETF flow, on the ETF premium. We present the estimates separately for creation days with a positive premium (Panel (a)) and redemption days with a discount (Panel (b)). In some specifications, we include the ETF return and NAV return on the day of the flow as controls. All regressions also include date and ETF fixed effects. The sample runs from May 2017 to February 2020, standard errors are clustered at the ETF and date levels, and t-stats are shown in parentheses. One, two, and three stars indicate statistical significance at the 10%, 5%, and 1% levels, respectively. See Section 4.5 for more details.