Will ETFs Drive Mutual Funds Extinct?*

Anna Helmke†

May 10, 2024

Click here for the latest version

Abstract

This paper challenges the conventional wisdom that exchange-traded funds (ETFs) are more liquid than open-end mutual funds. I build a model and establish that same-index ETFs and mutual funds provide liquidity at different horizons. Investors facing higher (lower) liquidity risk and thus shorter (longer) investment horizons prefer mutual funds (ETFs). Since they can be redeemed at NAV, mutual funds holding illiquid assets provide higher short-term liquidity, but the resulting payoff complementarities make them underperform ETFs in the long run. ETFs, however, are subject to mispricing and illiquidity in the short term due to arbitrageurs' balance-sheet constraints. In equilibrium, both funds coexist when investors face heterogeneous liquidity needs. The model generates novel, testable predictions concerning the competition and future trajectory of index ETFs and mutual funds.

Keywords: ETFs, index mutual funds, liquidity provision, portfolio choice

JEL Codes: G11, G23

*I thank Itay Goldstein, Olivia Mitchell, Luke Taylor, and Yao Zeng for providing me with invaluable guidance throughout the preparation of this work. I am grateful to Gonzalo Cisternas, Itamar Drechsler, Keshav Dogra, Thomas Eisenbach, Fulvia Fringuellotti, Anna Kovner, John Kuong (Discussant), Michael Lee, Antoine Martin, Steffen Mayer (Discussant), Darcy Pu, Nick Roussanov, Yujie Ruan (Discussant), Tom Sargent, Asani Sarkar, Sergey Sarkisyan, Or Shachar, Patrick Shultz, Huan Tang, Chaojun Wang, as well as seminar participants at the 20th Macro Finance Society Workshop, Finance Theory Group conference, Inter-Finance PhD seminar, TADC, Federal Reserve Bank of New York, EFA, INSEAD, IESE, Notre Dame, Vanderbilt, Colorado Boulder, Federal Reserve Board of Governors, Federal Reserve Bank of Boston and Wharton for helpful suggestions and conversations. I thank market participants at BlackRock and Citibank – Shreya Adiraju and Rudra Gopinath – for many helpful conversations. I gratefully acknowledge financial support from Analytics at Wharton, the Jacobs Levy Equity Management Center, Mack Institute for Innovation Management, and Rodney White Center for Financial Research. All errors are my own.

†Department of Finance, The Wharton School, University of Pennsylvania,
Email: at.helmke@gmail.com.

1 Introduction

A central idea in finance is that the average investor is best served by passively investing in a diversified index of securities. The growing adoption of this approach is evident in the shift from active to passive investing. Households generally rely on investment funds, such as open-end mutual funds (MFs), or exchange-traded funds (ETFs) to obtain their desired index exposure. MFs and ETFs both perform liquidity transformation by pooling money from many investors to purchase securities and hold virtually identical portfolios in the passive investment sector. Furthermore, they control nearly equal portions of the over $\$9tn$ in assets under management (AUM) within U.S-registered public index funds (figure A.1). ETFs, however, have consistently garnered greater inflows than their mutual fund counterparts for each of the preceding 15 years, leading some industry professionals to predict that mutual funds will become obsolete (figures A.2 and A.4). Nevertheless, index MFs have continued to receive net capital inflows. The gap between the perceived advantages of ETFs, such as superior tax efficiency in the U.S., intraday trading, and lower minimum investments, and the continued, substantial presence of index MFs begs the question what purpose MFs serve in the world of passive investing. Understanding the differences between index ETFs and MFs is important because of their size, central role in households' portfolio allocation, and the relevance to ongoing policy debates on regulatory measures to minimize the flow-induced dilution of shareholders' interest in open-end funds.

This paper seeks to rationalize the coexistence of same-index ETFs and MFs with their unique liquidity-provision services. I propose a static, discrete-time portfolio choice model featuring rational investors with ex-ante heterogeneous liquidity needs who allocate their wealth between an index ETF and MF. My primary contribution is to emphasize the unique liquidity-transformation characteristics inherent in alternative index fund structures. ETFs' and MFs' distinct trading and pricing mechanisms affect fund liquidity provision across different horizons. The central distinction between these two funds lies in the fact that relative mispricing in ETFs affects investors' returns in the short term. In contrast, MF share dilution negatively impacts investors' returns in the long term. I establish that the exchange-traded nature of ETFs does not inherently make them more liquid than MFs. Instead, excessive price fluctuations associated with ETFs' intraday trading reduce their short-term liquidity-provision compared to MFs. Rational index investors can optimize their portfolio allocation by selecting the fund type that aligns best with their specific liquidity requirements and investment horizon.

In my model, investors generally invest for the long-term but may receive liquidity shocks, necessitating the prompt liquidation of their portfolio holdings. ETFs and MFs are simply financial technologies that facilitate index-based investments by providing diversification and access to otherwise inaccessible market segments. Both funds passively track the same benchmark index. ETFs are exchange-traded intraday at the prevailing market price, just

like stocks, whereas MFs are purchased or redeemed directly from the fund sponsor at the end-of-day fund net asset value (NAV). I show that both funds are fundamentally different investment vehicles. The key distinguishing features between ETFs and MFs are their distinct payoff structures, specifically the mechanisms determining at what price their fund shares can be redeemed at short notice.

I derive three main findings. First, I establish that the payoff structure differences between index ETFs and MFs matter for fund liquidity provision at different horizons.[1] In this context, liquidity provision is defined by the relative payoffs of equity claims on identical index portfolios issued by different financial intermediaries (ETFs and MFs). ETFs provide more liquidity over the long-term, whereas MFs provide more liquidity at shorter horizons. Short term, within this framework, corresponds to a few trading days, whereas the long term extends across periods of months to years. Outflows from ETFs only have temporary effects on investors' payoffs, whereas MF flows have persistent effects on investors' payoffs.

These results emerge from the distinct mechanisms through which the costs of the liquidity transformation provided by these investment funds are reflected in their share prices. ETF liquidity provision occurs in secondary markets via authorized participants (APs). By contrast, MF liquidity provision takes place within primary markets among MF shareholders as funds themselves stand ready to sell or repurchase any quantity of fund shares at the end-of-day NAV on demand. Consequently, over the short-term, ETFs can be mispriced relative to their fund NAVs when balance sheet capacity constraints prevent APs from providing liquidity in secondary ETF markets via primary market creations or redemptions. Eventually, ETF prices converge back to the fund NAV once AP balance sheet capacity constraints subside. This theoretical result is consistent with the data. Throughout the March 2020 market sell-off, U.S.-based index ETFs on average traded at an discount of -32.5 bps (median discount of -8.3 bps) relative to their NAV. The mispricing was even larger in ETF market segments characterized by a greater liquidity mismatch. International equity ETFs experienced discounts exceeding 2%, while fixed income ETFs traded at discounts of approximately 3% on an asset-weighted basis during the peak crisis days.[2]

Unlike ETFs, MFs guarantee investors the ability to trade at the end-of-day fund NAV, regardless of prevailing financial market conditions. MFs' pricing mechanism gives rise to an externality between fund investors. While ETF prices can fluctuate excessively relative to the value of their underlying assets over the short term, MF NAVs can be insufficiently flexible as they do not fully reflect the transaction costs and price impacts associated with shareholders' redemptions. Accordingly, MF redemptions dilute the shareholdings of the remaining fund investors. This share dilution represents the liquidity premium investors pay in exchange for the short-term liquidity protection offered by MFs, should they choose not to redeem their

[1] To avoid ambiguity, MFs in my paper refers to index mutual funds, unless otherwise specified.

[2] See figures A.5 - A.20 for data on the relative mispricing in ETFs over time and across asset classes.

shares prematurely. Due to the in-kind nature of ETF creations and redemptions, there is no share dilution risk in ETFs.

I demonstrate that ETFs' market-based pricing mechanism gives rise to reverse run incentives, as strategic substitutabilities encourage shareholders to remain invested when intermediaries are balance sheet constrained. Investors who do not need immediate access to liquidity will always abstain from selling their ETF shares prematurely. The opposite is true for MFs. The insufficient flexibility of MF prices leads to payoff complementarities, encouraging early redemptions by long-term investors during periods of market illiquidity, potentially culminating in mutual fund runs.

Importantly, in my model, asset market illiquidity serves as the fundamental prerequisite for the inherent frictions within ETFs and MFs. In highly liquid index segments, such as large cap domestic equities (e.g., S&P 500 index funds), the relative mispricing risk in ETFs as well as the share dilution costs in MFs are small. I establish that in such cases, both, ETFs and MFs, offer virtually frictionless liquidity transformation. By contrast, in relatively more illiquid and increasingly popular market segments such as corporate bonds or international equities, the financial frictions tied to ETFs and MFs are predicted to attain economic significance, leading to evident divergences in their respective payoffs across investment horizons. In this context, financial frictions impede the seamless functioning of index fund markets.

In my second main finding, I show that within the realm of funds tracking benchmark indices composed of imperfectly liquid securities, funds' relative liquidity differences give rise to a cut-off equilibrium wherein investors self-select into ETFs versus MFs depending on their liquidity needs and expected investment horizons. Understanding funds' different pricing risks and run incentives, an investor with a long horizon ex-ante will optimally select the ETF, while an investor with a shorter horizon will select the MF. Investors with lower liquidity risks or longer horizons are positioned to circumvent the potential short-term mispricing of ETFs. By contrast, opting for MFs would likely expose them to the costs associated with the earlier redemptions by shorter-term fund investors. Conversely, when investors anticipate a need to quickly sell their investments within a few days, they highly value the immediate liquidity protection offered by MFs. These are agents who invest to hedge against liquidity needs in different states of the economy, for example due to higher labor income risk, unforeseen expense shocks, or the need to support their lifestyles through income from capital investments. They are willing to forgo long-term expected returns to sidestep the potential short-term mispricing in ETFs.

Third, I investigate the consequences of the coexistence of same-index ETFs and MFs for funds' vulnerabilities to outflows and for investors' payoffs. I demonstrate that the existence of ETFs makes MFs more vulnerable to premature investor redemptions, and the existence

of MFs reduces ETF mispricing. Investors who are likely to require urgent liquidity in the future are worse off after the introduction of an ETF compared to the MF-only equilibrium. When index fund assets are divided between an ETF and MF, the liquidity risks encountered by investors within the MF are pooled among a smaller set of agents, thereby diminishing the extent of liquidity co-insurance provided to each individual investor within that group. Investors with low expected liquidity needs benefit. The competition between ETFs and MFs allows them to separate from higher liquidity risk investors, thereby avoiding the transaction costs associated with those investors' short-term liquidity needs.

These insights can inform decision-making processes for investors, regulatory authorities, and asset managers. Despite the seemingly straightforward nature of index investing, investors should look beyond fees or taxes and remain cognizant of the potential indirect liquidity-provision costs tied to their chosen investment vehicle. This study also offers implications for index fund selection within retirement accounts, an area of paramount significance. To avoid a scenario where retirement investors in illiquid index MFs during the wealth accumulation phase inadvertently subsidize short-term liquidity provision to retirees and non-retirement account holders, retirement plan sponsors should add ETFs to the menu of investment options. An awareness of investors' trade-offs between ETFs and MFs can further assist regulators in designing fund liquidity management tools. In my model, MFs are subject to run risk. ETFs' payoff structure discourages early redemptions by patient investors but imposes excessive liquidation costs on impatient investors. One potential tool to reduce flow-induced share dilution, and therefore run risk in MFs is swing pricing. Swing pricing is a mechanism that adjusts MFs' NAVs to account for flow-induced transaction costs, thereby protecting remaining shareholders' interests. In November 2022, the SEC proposed to make swing pricing mandatory for most U.S. based open-end mutual funds, sparking vehement criticism from both asset managers and politicians.[3] Contrary to the arguments of these stakeholders, I show that swing pricing can reduce incentives for early MF liquidations while preserving MFs' relative appeal to investors with higher liquidity needs. If certain conditions are met, including optimal swing factor calibration, along with rational and forward-looking investors, swing pricing even enables MFs to dominate same-index ETFs in terms of liquidity provision. My paper further suggests that regulators should avoid endorsing multi-share class structures, where ETF and MF share classes coexist within a single fund portfolio, especially when the underlying index consists of less liquid securities. This arrangement tends to favor MF shareholders at the expense of ETF investors. From an asset management perspective, this paper sheds light on the recent trend involving MF-to-ETF conversions by suggesting that the benefits of these reorganizations are larger for MFs with less liquid assets and longer-term investors.

[3]For details on the U.S. Securities and Exchange Commission's (SEC) proposed swing pricing rule, refer to https://www.sec.gov/files/33-11130-fact-sheet.pdf.

Many asset managers and financial advisors promote ETFs as more liquid or easier to trade than their MF peers. This may lead some to think that ETFs are more suitable for investors with short-term liquidity needs. I challenge this notion, demonstrating that ETFs are typically better suited for investors with lower liquidity risks or longer investment horizons, especially when underlying portfolio holdings are illiquid. Previously, expense ratios and the tax implications associated with fund-level capital gains realizations have dominated discussions on the merits of ETFs versus MFs. In fact, fee and tax disparities alone are insufficient to explain the coexistence of same-index ETFs and MFs. Both fund types charge virtually identical fees on an asset-weighted basis within the most popular index segments (figures A.21 and A.22), and although U.S.-based ETFs allow deferring capital gains taxes until investors sell their fund shares, ETFs have experienced similar growth in other markets where they are subject to identical tax rules as MFs. Accordingly, this paper puts forth a novel, rational explanation, based on funds' universally different security design, for the trillions of U.S. dollars still allocated to index MFs.

1.1 Related literature

I contribute to the growing ETF literature by formalizing a unified framework to understand investors' trade-offs between ETFs and MFs. I show that index ETFs and MFs are imperfect substitutes because of differences in the nature of their liquidity provision.

Few other papers have explicitly considered investor trade-offs between ETFs and open-end mutual funds, and the few that do, generally agree that ETFs and MFs can coexist in equilibrium. They attribute the coexistence of both fund types to clientele effects resulting from different fee structures, ETFs' intraday liquidity provision (Agapova 2011) and tax efficiency (Moussawi, Shen, and Velthius 2022). Specifically, Agapova (2011) predicts that investors with higher liquidity needs or longer time horizons prefer ETFs because of their exchange-traded nature and lower fees, while short-term traders prefer MFs because of their commission-free nature. Figure A.21 illustrates that index MFs are now available at similar expense ratios as their ETF competitors. Besides, many U.S. brokerage firms offer commission-free trading for retail investors. I contribute to this literature by showing that, even abstracting from fee and tax differences, ETFs provide relatively larger payoffs over the long term due to the absence of share dilution risks. In related work, based on a model featuring ETFs that are frictionless and identical to the benchmark index, Huang and Guedj (2009) show that ETFs are more suitable when investors have more correlated liquidity shocks or underlying securities markets are less liquid. I endogenize the frictions in both ETFs and MFs and demonstrate that MFs can be more suitable for investors with short-term liquidity needs due to the potential for relative mispricing in ETFs. Investors with intraday trading needs, such as hedge funds who trade ETFs for speculative or hedging purposes, are outside of the scope of my paper, as they would never choose to invest in MFs in the first place.

My paper also adds to the literature on investment fund choice of organizational structure. The previous literature largely focuses on the choice between the open- and closed-end fund (CET) structure. While I focus on the liability-side competition between same-index ETFs and MFs, this literature has emphasized complementarities in open- and closed-end fund asset holdings. Deli and Varma (2002) empirically show that funds holding securities with lower liquidity or price transparency are more likely to be structured as CEFs. Similar to my model, Cherkes, Sagi, and Stanton (2009) theoretically argue that CEFs exist because they facilitate investments in illiquid securities without the externality costs of open-end MFs. Elton, Gruber, Blake, and Shachar (2013) highlight CEFs' ability to use leverage. I extend this line of research by considering the trade-off between the ETF and the open-end MF structure. ETFs represent a hybrid organizational form between open- and closed-end MFs. Their shares are exchange-traded, similar to CEFs, but they also incorporate an intermediary-based mechanism for the creation of new shares and the redemption of existing ones, akin to the structure found in open-end MFs.[4]

This paper further complements the literature on liquidity provision by non-bank financial intermediaries (NBFIs) and their associated risks. I build on work by Diamond and Dybvig (1983) on bank liquidity provision. In contrast to their framework, investors in my model are ex-ante heterogeneous and invest in index fund equity instead of bank deposits. Equity-issuing intermediaries including ETFs and MFs also provide liquidity (Ma, Xiao, and Zeng 2022a), and similar to banks, the liquidity mismatch between ETF or MF shares and their portfolio holdings is the key driver of frictions in this paper. For MFs, the literature has identified the combination of payoff complementarities for MF investors with the liquidity mismatch on MFs' balance sheet as the central source of run risk. Chen, Goldstein, and Jiang (2010) show empirically and theoretically in a global games framework that the guaranteed redemption at the fund NAV gives rise to a first-mover advantage among investors which increases in the illiquidity of fund assets. Payoff complementarities in MFs arise because redemptions are associated with fund costs not fully reflected in fund NAVs at which exiting investors trade. Edelen (1999) quantifies the significant cost of liquidity-motivated trading for long-term MF investors, and Coval and Stafford (2007) show that MFs tend to conduct costly and unprofitable trades ex-post large outflows at the cost of their remaining shareholders. They estimate that most flow-induced MF trades occur with a lag of one day after redemption events.[5] Dickson, Shoven, and Sialm (1999) provide evidence for the

[4]Another related literature strand studies competition among funds of the same type. Malamud (2016), Box, Davis, and Fuller (2019) and Khomyn, Putniņš, and Zoican (2023) examine the coexistence of same-index ETFs. Hortacsu and Syverson (2004), Elton, Gruber, and Busse (2004) and Choi, Laibson, and Madrian (2009) investigate the coexistence of same-index open-end MFs.

[5]See also Feroli, Kashyap, Schoenholtz, and Shin (2014) and Goldstein, Jiang, and Ng (2017). Falato, Goldstein, and Hortaçsu (2021) and Ma, Xiao, and Zeng (2022b) offer empirical evidence from the Covid-19 crisis. Kacperczyk and Schnabl (2013) and Schmidt, Timmermann, and Wermers (2016) document shareholder runs in money market funds. Previous attempts to quantify the costs of MF redemptions

importance of capital gains taxes as a source of externalities among MF investors. Most recently, the staleness in MF NAVs is documented by Choi, Kronlund, and Oh (2022) for fixed income funds; previous evidence focusing on international and illiquid domestic equity MFs include Goetzmann, Ivković, and Rouwenhorst (2001), Chalmers, Edelen, and Kadlec (2001), Boudoukh, Richardson, Subrahmanyam, and Withelaw (2002) and Zitzewitz (2003). Another source of staleness in MF NAVs are outdated portfolio weights (Tufano, Quinn, and Taliaferro 2012). I contribute to this literature by showing that the lag with which flow-induced transaction costs are reflected in MF NAVs provides liquidity insurance to investors with urgent liquidity needs. Yet, over the longer term this liquidity insurance comes at the cost of share dilution for the remaining MF investors. Consistent with prior studies, these payoff complementarities among MFs investors introduce MF run risk into my model.

The literature on risks in ETFs has focused on the spillover effects from ETFs to financial markets caused by the AP arbitrage channel.[6] From an investor's perspective, the key friction in ETFs is the potential for relative mispricing between the ETF price, at which investors can trade, and the fund NAV, which is intended to reflect the fundamental value of a share in the fund. Empirically, Haddad, Moreira, and Muir (2021) document large discounts in bond ETFs during the Covid-19 sell-off. Petajisto (2017) shows that the relative ETF mispricing remains statistically and economically significant, even after accounting for potential staleness in fund NAVs. Other related studies include Todorov (2021), Gorbatikov and Sikorskaya (2022) and Malamud (2016). Pan and Zeng (2019) demonstrate how the liquidity mismatch between ETF shares and portfolio securities, coupled with APs' balance sheet constraints, leads to limits to arbitrage in ETF markets. Shim and Todorov (2022) suggest that ETFs may be more effective in managing illiquid assets compared to MFs because APs can limit adverse spillover effects from fund liquidations to asset markets. I add to this literature by demonstrating that ETFs' distinct payoff structure discourages fund liquidations when markets are illiquid and endogenously attracts relatively more long-term investors.

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3 presents the equilibrium predictions. Section 4 analyzes policy implications. Section 5 discusses empirical predictions. Section 6 concludes.

include Chordia (1996), Wermers (2000), Greene and Hodges (2002) and Johnson (2004).

[6]For example, previous studies have analyzed the effect of ETF ownership on non-fundamental asset price volatility (Dannhauser and Hoseinzade 2017; Ben-David, Franzoni, and Moussawi 2018), asset price discovery (Brown, Davies, and Ringgenberg 2021; Israeli, Lee, and Sridharan 2017; Glosten, Nallareddy, and Zou 2021; Madhavan and Sobczyk 2016; Bhattacharya and O'Hara 2018) and return co-movement (Da and Shive 2018; Shim 2019; Madhavan and Morillo 2018).

2 Theoretical framework

I propose a portfolio choice model to study the tradeoffs faced by rational investors with heterogenous liquidity risks when choosing to allocate their portfolio between index ETFs and index open-end mutual funds built upon Diamond and Dybvig (1983). Investors generally invest for the long term, but they may occasionally need to liquidate financial assets at short notice in times when overall market liquidity is scarce. The model is designed to mimic adverse states of the economy characterized by low asset market liquidity and high demand for liquidity. Asset market liquidity refers to the ease with which financial securities can be traded. Correspondingly, liquidity risk denotes the probability that an investor needs to liquidate her financial assets for consumption purposes at times when it is costly to do so. On the one hand, these episodes of market stress represent the states in which the distinct frictions associated with same-index ETFs and MFs materialize and meaningfully impact fund liquidity provision. On the other hand, it is during adverse market conditions that funds' liquidity-provision service is most valuable for investors.

I focus on investors' discretionary portfolio allocations outside of retirement savings accounts. This allows me to capture a wide range of investment motives and liquidity needs, crucial for examining liquidity-provision differences across competing investment funds. Implications for investments within retirement accounts are discussed in section 4.2. More generally, this framework is not limited to retail investors. It is sufficiently flexible to accommodate any type of investor relying on investment funds.

The model is designed to capture the fundamental economic differences between ETFs and MFs, the mechanisms by which both funds' shares are priced and traded in financial markets. I abstract from variation in fund expense ratios and capital gains taxation and refer to Agapova (2011) and Moussawi et al. (2022) for an empirical analysis of the implications of fee and tax differences between ETFs and MFs. While fees and tax rules are at the discretion of fund sponsors and regulators of the fund's domicile country, the pricing and trading mechanisms constitute intrinsic components of the structures of both ETFs and MFs. There are three periods, $t = \{1, 2, 3\}$; time is discrete and financial markets are competitive. There is a single consumption good, dollars. The economy consists of investors i and of financial intermediaries. Investors are heterogeneous in terms of their liquidity risk exposure. Liquidity risk in this context refers to the possibility that unexpected financial needs or obligations may arise, requiring investors to sell their investments quickly under less-than-ideal market conditions. Some investors are more likely to be forced to liquidate their assets earlier than others. There is also a group of "sleepy" or inattentive investors; for the ease of exposition, I call them retirement investors. Consistent with the menu of funds in practice offered by many retirement plans, retirement investors can only invest in mutual funds. Generally, these sleepy investors reflect agents who do not actively monitor or adjust their investment portfolios in response to changing market conditions or new information. They follow a simple

buy-and-hold strategy, making no portfolio re-allocation decisions between the initial and terminal period.

Financial markets feature risky composite securities: index MFs, and index ETFs both of which are investment technologies that pool money from investors to invest in a diversified portfolio of potentially illiquid securities. For tractability, I abstract from individual securities markets; ETFs and MFs simply hold the composite security. The composite security represents a benchmark index and can only be traded by fund managers and intermediaries. Investors can only invest in composite securities via investment funds, and they cannot trade directly in the index. There is also a risk-free asset which pays zero interest $R^f = 1$.

Key model frictions directly arise from the different payoff structures of ETFs and MFs. They emerge in ETFs from the potential for the prices at which investors transact in secondary markets to diverge from the fund's net asset value. I refer to this difference between ETF market prices and fund NAV as relative mispricing. If the ETF trades at a discount (premium), long ETF investors receive a payoff below (above) the funds' fundamental value. In contrast, frictions in MFs are due to payoff externalities between fund investors. There are no agency frictions between fund sponsors, intermediaries and investors. Objectives of fund sponsors are outside of the scope of my model.

Figure 1: Timeline of information and investor actions

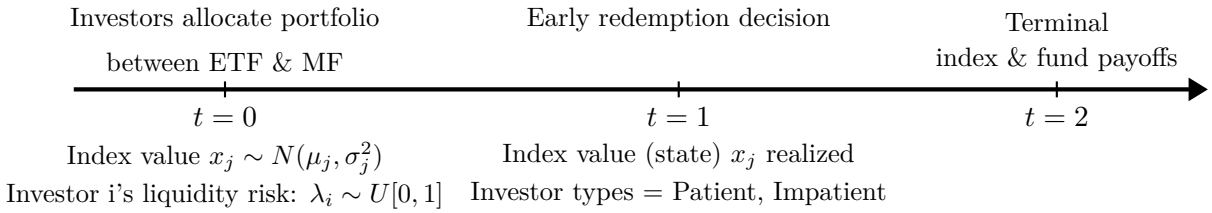


Figure 1 summarizes the model information structure and the timing of agents' decisions. The terminal index payoff, x_j, serves as the state variable, reflecting the exogenous fundamental value of the composite security j, is distributed $x_j \sim N(\mu_j, \sigma_j^2)$ with $\mu_j > 1$ and observed at the beginning of period $t = 1$. $\mu_j < 1$ implies a positive expected index return. I focus on investors' portfolio allocation decision between the ETF and MF within a given benchmark segment j. Since my focus is on the tradeoffs across index fund types, I do not consider investors' portfolio choices across different asset classes and index segments. The relation between the fundamentals x_j across j is not relevant for this paper's analysis. All agents have identical information about x_j. Table D.1 summarizes notation.

The model is intentionally designed to capture adverse economic conditions characterized by an overall net demand for liquidity among fund investors, manifested through redemptions

from MFs and sales of ETFs. In the interim period, individual investors confront two possible scenarios: a liquidity shock necessitating the immediate liquidation of their entire portfolios, or the absence of a liquidity shock allowing them to choose between immediate or future portfolio liquidation. Investors can be impatient consumers but not impatient savers. Given the lack of positive cash flows shocks, there are no fund inflows at $t = 1$. There is no potential for trade among fund investors and Jacklin (1983)'s critique does not apply. In fact, following the assumptions placed on the distribution of investor liquidity risks, outflows from the fund sector are strictly positive at $t = 1$. Even though liquidity risks are independent at the investor level, the model features systemic liquidity risks at the aggregate level. At least half of fund investors will liquidate their portfolio holdings early. As a result, investors are unable to fully insure themselves against liquidity risks: they depend on financial markets and intermediaries for liquidity provision. In practice, even during periods of market stress, index funds typically experience inflows from some investors who consistently allocate a fixed dollar amount to these funds as part of a savings plan. In this model, fund outflows can be understood as investors' net withdrawals.

2.1 Financial markets

Benchmark index. A benchmark index j represents a diversified portfolio of securities (e.g., S&P 500 Index, Bloomberg Aggregate Bond Index, MSCI Emerging Markets Index). In what follows, I focus on a single index j tracked by one ETF and one MF. The terms composite security and index are used interchangeably. It can be traded by mutual fund managers, financial intermediaries in their role as APs, and market makers (broker-dealers) specialized in providing liquidity in these index markets at $t = 1$. Composite security markets are segmented, a standard assumption in the literature (e.g., Malamud (2016), Gromb and Vayanos (2002)) which allows for potential relative mispricing between markets. This implies that MFs and APs trade with distinct index market makers, potentially at different prices. It is motivated by the different timing of when MFs and APs trade in index markets: APs respond to demand imbalances in secondary ETF markets in real time intraday, while MFs adjust their portfolio holdings with a lag after observing their investors' net redemptions. More fundamentally, the segmented index markets assumption allows distinguishing ETF and MF respective price impacts in the underlying security markets. MFs and APs are both price takers and do not compete on the asset side in index markets. The index price in the interim period is endogenous, given by P_t^j and determined by market clearing between market makers, APs and MFs. In the terminal period, the index pays a terminal dividend equal to its fundamental value, $Div_2^j = x_j$. Hence, the cum-dividend index price at $t = 2$ is $P_2^j = x_j$.

In an extension of the model, I relax the segmented market assumption and show that the same overall predictions continue to hold when ETFs and MFs trade with a common index

market maker.[7] The extension provides additional insights on how asset-side competition between ETFs and MFs affects fund investors' payoffs. Investors cannot directly trade composite securities. In theory, it is equivalent to assuming that the transaction costs investors face under autarky when trying to replicate, and sequentially rebalance, the index themselves, for instance through direct indexing in separately managed accounts (SMAs), are always strictly larger than the cost of holding an ETF or MF.

Investment funds. There exists both one open-end mutual fund as well as one ETF passively tracking every benchmark, j. Funds' ex-ante choice about which benchmark index to track is outside the scope of this model. I consider the case of a single benchmark index so the notation j illustrates the dependence of key parameters, and consequently the model predictions, on the return distribution and liquidity of the benchmark portfolio. Hence, j generally denotes the overall liquidity of the asset market segment under consideration. For simplicity, neither ETFs nor MFs charge fees, so management expense ratios are zero, $f_t^{ETF,j} = f_t^{MF,j} = 0 \; \forall t$ and j. Funds hold no cash, so the portfolio holdings of the ETF and MF tracking benchmark j are identical. Abstracting from portfolio differences allows me to focus on non-portfolio differences as the source of heterogeneity between ETFs and MFs.

ETFs. The ETF market price, $P_t^{E,j}$, arises endogenously intraday in equilibrium from market clearing between ETF investors and APs.

Assumption 1 *ETF creation and redemption baskets are identical to the benchmark index.*

Lemma 1 *The ETF NAV is always equal to the value of the benchmark index,* $NAV_t^{E,j} \equiv P_t^j$. *The ETF perfectly replicates its benchmark. There is no tracking difference.*

Here, tracking difference refers to the difference between the fund NAV and the index price, $NAV_t^{E,j} - P_t^j$. In practice, however, tracking difference often refers to the return difference between a fund and its benchmark. ETF returns and payoffs are equivalent in this model.

What matters to ETF investors is not the ETF NAV but the ETF market price at which they trade. $P_t^{E,j}$ is linked to the index price, P_t^j, via arbitrage. There are limits to arbitrage, so the ETF price can deviate from its NAV, resulting in relative mispricing vis-à-vis the benchmark index. Arbitrage constraints arise endogenously from APs' balance sheet capacity constraints.

Definition 1 *The relative ETF mispricing refers to the difference between the ETF's net asset value,* $NAV_t^{E,j}$, *and its market price,* P_t^E. *At any given point in time, the relative mispricing is given by:*

$$\epsilon_t^{E,j} \equiv NAV_t^{E,j} - P_t^E. \tag{1}$$

[7] The assumption that market makers are distinct from banks acting as APs in ETF markets is a simplification. In practice, different trading desks of the same broker-dealer may act in multiple roles as market maker in index markets, as a counterparty to MFs, as well as an AP for ETFs. Studying the implications of the associated conflicts of interests is beyond the scope of this paper but will be explored in future research.

When $\epsilon_t^{E,j} > 0$ ($\epsilon_t^{E,j} < 0$) the ETF trades at a discount (premium).

Then, the ETF price is given by $P_t^{E,j} = NAV_t^{E,j} + \epsilon_t^{E,j}$.

Given P_t^j, the mispricing captures the wedge between the index and the ETF price. Because transaction costs for trading ETF shares are incurred at the investor level, investors' effective (pre-tax) transaction price may deviate from $P_t^{E,j}$ by the amount of bid-ask spreads and trading commissions charged per ETF share. For simplicity, I assume ETF bid-ask spreads are zero, and I also assume investors trade ETFs commission-free.

Open-end mutual funds. All MF creation or redemption orders submitted throughout the trading day are executed at the end of each trading day, at the fund NAV. The daily MF NAV is equal to the closing index price. By design, MFs have zero relative mispricing:

$$P_t^{M,j} \equiv NAV_t^{M,j}.$$

Rational investors have perfect foresight and correctly anticipate the fund NAV, $P_t^{M,j}$, when submitting their MF orders throughout the trading day. This assumption is standard in the literature. The MF NAV is flexible, to the extent that it accounts for new information on the fundamental index value, x_j. Yet, the NAV is not fully forward-looking: total net fund flows and the price impact caused by MFs' ensuing flow-induced trading are generally unknown until the next trading day (Ma et al. 2022b). While ETF prices are determined in equilibrium from the intraday supply-demand conditions in ETF markets, MF prices are set based on supply-demand conditions in index markets at the end of the trading day. The imperfect flexibility of MF prices constitutes the key friction of mutual funds in this model. Formally, this is captured by the specification of the MF NAV at $t = 1$ which is quasi-exogenous and given by:

$$NAV_1^{M,j} = \psi E_1[x_j]. \tag{2}$$

$P_t^{M,j}$ is a function of the expected index value and a parameter ψ with $0 < \psi \leq 1$. Through its dependence on x_j, the MF NAV reflects the latest information on the value of its asset holdings, where ψ reflects a discount or penalty for early liquidation of MF shares. The lower ψ, the higher the cost to the individual investor of liquidating her MF holdings early. $\psi < 1$ is a technical restriction. It is necessary to prevent early liquidation of all MF holdings, regardless of the index fundamentals, from becoming the sole dominant strategy for risk-averse investors. Intuitively, in the absence of any MF outflows at $t = 1$, investors can expect to earn a strictly positive return from holding the MF between the interim and terminal period. The model predictions are robust to alternative specifications as long as $NAV_1^{M,j} < x_j$ holds. In an extension in section 3.3.1, I fully endogenize the MF NAV at $t = 1$.

2.2 Agents

There is a continuum of individual investors with mass 1 and sleepy investors with mass η.

Individual investors. Individual investors are the agents of interest, broadly defined to include retail and institutional investors, such as endowments or family offices who may choose to invest in ETFs or MFs. In the U.S., around $68.6m$ (52%) and $16.1m$ (12%) of households invest in MFs and ETFs respectively (ICI 2023). Investors, such as high-frequency trading firms, who use ETFs for short-term bets or hedging purposes and require intraday liquidity are outside of the scope of the model. Since only ETFs provide intraday liquidity and MFs generally resrict investors' ability to frequently move in and out of a single fund, such agents would never invest in MFs in the first place. They do not face an interesting trade-off between different fund types.

The unit mass of investors is homogenous in terms of initial wealth, which I normalize to one unit of capital, $\theta_0^i = 1$. They invest over the long term but may at times face short-term liquidity shocks in the spirit of Diamond and Dybvig (1983). Liquidity shocks force investors to liquidate their assets at $t = 1$. In contrast to classic bank run models, in which all investors initially face the liquidity risk, here the ex-ante probability of a liquidity shock, denoted by λ_i, differs across investors. λ_i is independently and identically distributed according to:

$$\lambda_i \sim U[\lambda_0, \lambda_1],$$

where $0 \leq \lambda_0 \leq \lambda_1 \leq 1$ reflect the distribution of liquidity risk across agents in the economy. I assume $\lambda_0 = 0$ and $\lambda_1 = 1$: λ_i is privately observed at $t = 0$ and causes ex-ante heterogeneity among investors.[8] λ_i is directly linked to i's investment horizon, T_i, according to:

$$E_0[T_i] = 2 - \lambda_i.$$

The larger λ_i, the shorter i's expected investment horizon. Investors of type $\lambda_i = 1$ must always sell all assets at $t = 1$: These are investors who need liquidity on a specific date in the future. Type $\lambda_i = 0$ investors can wait until $t = 2$ before liquidating their portfolio holdings. I denote the expected mass of impatient investors, that is investors who receive a liquidity shock at $t = 1$, by $\bar{\lambda} = \int_i \lambda_i di = \frac{\lambda_0 + \lambda_1}{2}$. Under my baseline assumptions, $\bar{\lambda} = \frac{1}{2}$.

Heterogeneity in λ_i can reflect differences in investors' probabilities of being hit by an unexpected expense or income shock as well as differences in their holding periods. According to survey estimates, 26% of mutual-fund owning households use these investment vehicles as a tool to save for emergencies (ICI 2023). Intuitively, investors may face unforeseen costs such

[8] By varying λ_0 and λ_1, the model can be used to analyze the effects of changes in the distribution of liquidity risks in the economy on the allocation of assets between ETFs and MFs.

as unexpected medical expenses, home repairs (e.g., after a natural disaster), or sudden job loss, necessitating immediate access to liquidity beyond what they hold in cash and equivalents.[9] If liquid asset reserves are already exhausted due to unforeseen expenses, investors may further be compelled to sell fund holdings in order to meet mortgage payments, tuition fees, or other debt obligations. The sources of heterogeneity in λ_i are multifaceted. Households with higher incomes and lower leverage generally possess greater disposable cash flows, enabling them to cover unexpected expenses without liquidating assets. Insurance coverage and access to credit can also hedge against unforeseen health or property-related expenses. Differences in occupations can significantly impact job security, particularly during economic downturns. Households with more dependents or caregiving responsibilities may face heightened liquidity risks due to more frequent and diverse financial needs. Finally, investors early in their life cycles often maintain longer investment horizons, whereas those in retirement may depend on the periodic liquidation of financial assets to cover essential living expenses.

Given λ_i and $x_j \sim N(\mu_j, \sigma_j^2)$, at $t = 0$ investors choose to allocate their endowments between ETFs and MFs. They do not have access to the risk-free asset in the initial period, but instead they must invest their entire capital endowment in index funds. This restriction allows me to isolate the effect of different liquidity risk exposures on investors' allocations across fund types. After observing their idiosyncratic liquidity shocks, investors at $t = 1$ can decide to liquidate some of their fund holdings early and store the proceeds in the risk-free asset until the terminal period. When hit by a liquidity shock, they always liquidate their entire fund portfolios. Investors cannot reallocate assets between MF and ETF markets at $t = 1$. Conditional on their initial allocations to MFs and ETFs, $\{\theta^{i,M,j}, \theta^{i,E,j}\}$, they can only retain or liquidate fund shares, and they receive no income from sources other than their financial investments. They do not have access to leverage and are subject to short-selling constraints.

All investors have identical time-separable preferences. The primitive utility function, denoted as $u(c)$, is defined with respect to consumption. I assume investors are risk-neutral. In the absence of storage technologies for the consumption good cash, other than investment funds, an agent's consumption always equals her wealth at the end of her lifespan T_i, $c_T^i = w_T^i$. The investor chooses her initial portfolio allocation to maximize her expected utility over terminal wealth, represented as $E[u(w_T^i)]$, subject to her endowment, leverage, and short-selling constraints in equations 4 - 6 respectively:

[9] According to the Fed's 2022 Survey of Household Economics and Decisionmaking (SHED), 23 percent of adults had unexpected medical expenses in the prior 12 months, with the median amount between $\$1,000$ and $\$1,999$. During the same period, 13 percent of adults were directly affected by a natural disaster and 3 in 10 adults experienced income variability from month to month. See https://www.federalreserve.gov/publications/report-economic-well-being-us-households.htm.

$$\max_{\theta^{i,M,j},\theta^{i,E,j}} E[w_T^i] \tag{3}$$

$$s.t.\ \theta_0^{i,E,j} P_0^{E,j} + \theta_0^{i,M,j} P_0^{M,j} = 1, \tag{4}$$

$$\theta_0^{i,E,j} - \theta_1^{i,E,j} > 0,\ \theta_0^{i,M,j} - \theta_1^{i,M,j} > 0, \tag{5}$$

$$\theta_t^{i,E,j} \geq 0,\ \theta_t^{i,M,j} \geq 0\ \forall\ t = 0, 1. \tag{6}$$

$\theta_0^{i,E,j}$ and $\theta_0^{i,M,j}$ are the number of shares of the ETF and MF tracking index j bought by investor i at $t = 0$, respectively. The index serves as the numeraire in the economy.

The terminal wealth of patient or late investors, characterized by $T_i = 2$ and denoted by $i = l$, is given by:

$$w_2^i = \theta_1^{i,E,j} P_2^{E,j} + \theta_1^{i,M,j} P_2^M + (\theta_0^{i,E,j} - \theta_1^{i,E,j}) P_1^{E,j} + (\theta_0^{i,M,j} - \theta_1^{i,M,j}) P_1^{M,j} \quad \forall\ i = l. \tag{7}$$

Investors subject to liquidity shocks in the interim period liquidate all portfolio holdings at $t = 1$. At that time, these impatient or early investors are characterized by their investment horizon $T_i = 1$ and denoted by $i = e$. They derive utility only from their wealth in the interim period, w_1^i. The terminal wealth of impatient investors is given by:

$$w_1^i = \theta_0^{i,E,j} P_1^{E,j} + \theta_0^{i,M,j} P_1^{M,j} \quad \forall\ i = e.$$

The terminal wealth of impatient investors depends solely on their initial allocations, $\theta_0^{i,E,j}$ and $\theta_0^{i,M,j}$, as well as on the fund prices at $t = 1$.

Sleepy MF investors. Sleepy investors' portfolio allocations are exogenous, and they invest their entire endowment in MFs. There is a mass η of them who exist for technical reasons to maintain model tractability. They prevent MFs from going bankrupt in scenarios where MF runs occur in equilibrium. This simplifies the model solution and is consistent with modeling conventions in the previous MF literature (Chen et al. 2010). Sleepy investors are not exposed to short-term liquidity risks: they never rebalance their portfolio holdings during the interim period and do not engage in early withdrawals. Instead, they hold their fund shares until maturity. The terminal value of their aggregate fund portfolio value is given by:

$$W_2^{Sleepy} = \eta P_2^{M,j}.$$

Without loss of generality, the model can be extended to also include sleepy ETF investors. Their inclusion neither alters the model's predictions nor simplifies its solution. Consequently, sleepy ETF investors are omitted for simplicity in the discussion below.

2.3 Financial intermediaries

Financial intermediation in the economy occurs at two levels: in fund and index markets. In fund markets, the representative authorized participant (AP) and mutual fund provide liquidity to their investors, while in index markets the AP and MF demand liquidity from index market makers who act as the central index liquidity supplier.

Index market makers. Risk-neutral index market makers are the broker-dealers in composite security markets: they appear only in the interim period, and their sole purpose is to provide liquidity to investment funds. These intermediaries are not frictionless: broker-dealers need to hold and manage a stock of potentially illiquid index constituent securities, incurring capital charges and funding costs, which can erode their profit margins and constrain their market-making activities. Formally, I model these frictions as quadratic inventory costs. In equilibrium, index market makers absorb the supply of index shares resulting from MFs' flow-induced portfolio liquidations and AP arbitrage activities. ETF arbitrage occurs intraday at $t = 1$, while MFs trade in index markets with a lag on the next trading day at $t = 1^+$, after observing their investors' net redemptions. To match this institutional feature of fund markets, I assume that index markets are segmented: APs and MFs trade with distinct but identical index market makers. This convention allows me to focus on liability-side competition between MFs and ETFs. Each representative market maker submits a price schedule for index shares, $P_t^j(\Theta_t^{D,j})$, to maximize her expected trading profits given her inventory costs. Market makers' optimization problem is given by:

$$\begin{aligned} &\max_{\Theta_t^{D,j}} E_t[\Pi_{t+1}^j] \\ &s.t.\ \Pi_{t+1}^j = P_{t+1}^j - P_t^j - \frac{c_j}{2}(\Theta_t^{D,j})^2 \quad \forall\ t = 1, 1^+. \end{aligned} \tag{8}$$

For each market maker, equation 8 implies a downward sloping index demand schedule:

$$P_t^j = E_t[P_{t+1}^j] - c_j \Theta_t^{D,j} \quad \forall\ t = 1, 1^+. \tag{9}$$

P_t^j is the index price offered by dealer D for $\Theta_1^{D,j}$ units of index shares at t.

The inventory cost parameter, c_j, is increasing in the index segment j specific liquidity. Intuitively, $c_{Equity} < c_{Corporate\ Bond}$: the inventory costs of holding corporate bonds are strictly larger than those of holding stocks. Intuitively, it generally takes the market maker longer to find a buyer for less liquid securities. In the meantime, the securities remain on the market maker's balance sheet and prevent it from using the capacity for market making purposes. Market makers pass on this cost to APs and MFs. When $c_j > 0$, MFs and APs have price impact when trading in index markets. This represents the baseline specification. Generally, c may be a function of both, security market j and aggregate market liquidity conditions.

This is consistent with Ma et al. (2022b) who, amongst others, find that MFs had a significant price impact even in generally liquid market segments during the Covid-19-related market sell-off. Since my model is conditional on an adverse state of the economy, this specification is not necessary. For $c_j = 0$, the model nests the case in which funds have zero price impact in security markets, a condition that may be satisfied for large cap domestic equities during periods when financial markets overall are liquid.

The representative market maker faces identical inventory costs each period, at $t = 1$ and $t = 1^+$, so APs and MFs face the same net demand for index shares in composite security markets. Index market price impact generated by the AP intraday does not impact the market conditions faced by MFs on the next trading day. In section 3.3.1, I extend the model to also allow or asset-side competition between funds. In this case, transactions in index markets by APs and MFs are interconnected because APs and MFs trade sequentially with the same representative market maker. First, the AP trades with the index market maker intraday at $t = 1$ to offload index shares obtained as a result of ETF redemptions. Next, the MF trades with the index market maker at the already depressed prices to satisfy its investors' net redemption requests. Thereby, the AP's price impact affects the liquidity conditions faced by the MF. The index demand schedule faced by the AP is still given by equation 9, and the index demand schedule faced by the MF is given by:

$$P^j_{1^+} = P^j_1 - c_j \Theta^{D,j}_{1^+},$$

where P^j_1 follows from equation 9 under market clearing between the market maker and AP.

Authorized participants. APs are deep pocketed risk-neutral financial intermediaries with the right to create and redeem ETF shares outright, via in-kind transactions with the ETF sponsor. This role turns APs into the central and only counterparty between ETF investors and fund sponsors. If liquidity shocks or optimal portfolio reallocation decisions by investors at $t = 1$ generate excess demand (supply) for ETF shares, the AP can step in and create (redeem) ETF shares by buying (short selling) the underlying creation (redemption) basket in index markets and delivering (redeeming) the proceeds (corresponding security basket). APs trade in index and ETF markets with the purpose of generating short-term profits. In practice, there are also other reasons for APs to engage in ETF creations or redemptions: for example, ETF arbitrage can help financial intermediaries manage their own liquidity risks or hedge balance sheet exposures. I abstract from such motives, so APs immediately offload any index or ETF positions obtained as part of their arbitrage trades in the same period $t = 1$. They do not hold any assets overnight. There exists a single representative AP for each ETF j. Like index market makers, APs only operate in the interim period, $t = 1$.

There are also costs associated with the AP's ETF arbitrage trades. In their role as broker-dealers APs face regulatory capital constraints. Even though in my model ETF arbitrage

trades are completed within period $t = 1$, the potential time lag between an AP's transactions in ETF markets and its offsetting trade in composite security markets implies that ETF arbitrage is not risk-free. Regulatory capital requirements applied to temporary index or ETF balance sheet positions give rise to limits to arbitrage. Pan and Zeng (2019) document that AP balance sheet capacity constraints distort AP arbitrage of corporate bond ETFs. In addition, when ETF track indices composed of illiquid securities that are traded in over-the-counter markets, such as corporate bonds, ETF arbitrage entails search and matching costs (Koont, Ma, Pastor, and Zeng 2023). Finally, hedging ETF creation or redemptions using derivatives is also costly.

Because regulatory capital requirements generally depend on intermediaries' balance sheet size, AP arbitrage costs increase in the size of ETF creations or redemptions and are thus modeled as variable costs.[10] In this model, balance sheet costs are symmetric for both ETF creations and redemptions. Yet, the specification of investors' liquidity shocks implies that APs solely redeem ETF shares in exchange for index shares in equilibrium. Accordingly, balance sheet costs can be interpreted to reflect the costs associated with ETF redemptions.

The AP's profit maximization problem at $t = 1$ is:

$$\max_{\Theta_1^{E,AP,j}} \Pi_1^{AP}(\Theta_1^{E,AP,j}) \tag{10}$$

$$s.t.\ \Pi_1^{AP} = (P_1^j - P_1^{E,j})\Theta_1^{E,AP,j} - \frac{1}{2}\phi_j\ (\Theta_1^{E,AP,j})^2. \tag{11}$$

Equation 11 is the AP's profit function. The AP never engages in any ETF trades at a loss. P_1^j is the index price and $P_1^{E,j}$ is the ETF price per share. The AP takes into account price impact in index markets when deciding on its optimal arbitrage strategy. Given the equilibrium index price, $P_1^{E,j}$ will be the result of the AP's constrained optimal arbitrage activity in ETF markets. $\Theta_1^{E,AP}$ is number of ETF shares redeemed by the AP. $\Theta_1^{E,AP} < 0$ implies that the AP creates $\Theta_1^{E,AP}$ units of new ETF shares by delivering a basket of $\Theta_1^{E,AP}$ units of the composite security to the ETF sponsor. For $\Theta_1^{E,AP} > 0$, the AP redeems $\Theta_1^{E,AP}$ existing ETF shares and receives a basket of $\Theta_1^{E,AP}$ units of the composite security from the ETF sponsor in exchange.

$\phi_j \geq 0$ is the balance sheet capacity parameter that captures variable transaction costs and balance sheet risks associated with ETF arbitrage. In periods of reduced market liquidity, as depicted in this model, APs often encounter more stringent balance sheet capacity constraints due to decreased asset valuations, increased loan loss provisions, and elevated refinancing expenses, which curtail their capacity to support ETF arbitrage transactions. In equilibrium,

[10] There are also fixed costs of ETF creations and redemptions. ETF sponsors charge a fixed fee per creation (redemption) unit to cover administrative costs. Yet, these costs tend to be small economically relative to the AP's trading related costs.

$\phi_j > 0$ generates limits to arbitrage. ϕ_j is a function of the unconditional index segment j specific liquidity. The time it takes them to liquidate index shares obtained as a result of an ETF redemption trade, depends on the index specific market liquidity. Due to the over-the-counter nature of corporate bond markets and the potential time lag between the market hours in domestic ETF and international equity markets, redemptions of corporate bond or international equity ETFs generally consume more balance sheet capacity than redemptions of large cap equity ETFs which can be offloaded from APs' balance sheet to liquid securities market quickly without significant price impact. The dependence of ϕ on j captures this feature of AP arbitrage: $\phi_{Corporate\ Bond} > \phi_{Equity}$. Overall, the specification of ϕ_j gives rise to cross-sectional variation in AP arbitrage across ETF market segments.

Assumption 2 *Over the long term, at $t = 2$, APs face no balance sheet capacity constraints. They resume ETF arbitrage until the ETF and index price have converged. Formally:*

$$\epsilon_2^{E,j} = 0.$$

Open-end mutual funds. Open-end mutual funds pool capital from investors to purchase securities. Due to the passive nature of index funds, MFs do not have any discretion over portfolio allocation decisions. All trades in index markets by MFs are flow-induced. I abstract from fund management fees, so there is also no scope for profit maximization. While ETF liquidity provision is market-based and delegated to the AP, MF companies directly supply liquidity to their investors. Whenever the MF experiences net capital inflows, it will purchase index shares of equivalent value in securities markets. Amid net capital outflows, the MF offloads the amount of shares required to repay redeeming investors at the fund NAV taking as given the index demand schedule submitted by competitive market makers. All net MF redemptions are executed in cash at the end of trading day $t = 1$, and MFs cannot satisfy redemptions using in-kind transfers (RIK) of security baskets.

In any period, t, the quantity of index shares sold by the MF j is given by:

$$\Theta_{t^+}^{M,j} = \frac{\Delta X_t^{M,j} P_t^{M,j}}{P_{t^+}^j}, \tag{12}$$

where the net fund flows are defined as the product of the fund NAV, $P_t^{M,j}$, and the change in the number of fund shares outstanding between the end of time t and $t-1$, represented as $\Delta X_t^{M,j} = X_{t-1}^M - X_t^{M,j}$. If $\Delta X_t^{M,j} > 0$, the MF experiences net redemptions at t. The index price faced by the MF, $P_{t^+}^j$, is endogenously determined by market clearing with their representative index market maker at $t = 1^+$.

While investors' net redemptions and the fund NAV are determined by the end of period t, due to the specific structure of U.S. MF markets $\Delta X_t^{M,j}$ is generally unknown to the fund until after the conclusion of the trading day or even the commencement of the following

day. Equation 12 captures this lag in the transmission of information on investor trades to the MF. In order to trade in index markets, MFs require accurate information regarding the volume of fund subscriptions or redemptions. In alignment with these inherent institutional constraints, the subscript t^+ denotes the beginning of the subsequent trading day following t. Following the aggregation of all creation and redemption orders, at t^+ the MF proceeds to sell (or buy) index shares to satisfy net redemptions (or fulfill net creations).

In practice, index MF managers could use cash to actively manage fund redemptions: they may first meet net redemptions by depleting cash holdings and only later sell portfolio securities to restore fund liquidity buffers. The evidence regarding the effectiveness of cash buffers in reducing MF fragilities is mixed.[11] I therefore abstract from such MF liquidity management. This simplification, however, introduces a gap between the time when MFs must pay off redeeming investors in cash (at $t = 1$), and when MFs receive the cash proceeds from their trades in index markets (at $t = 1^+$). To overcome this technical inconsistency, in the model I posit that MFs initially meet net redemptions at $t = 1$ by drawing down overnight credit lines from exogenous banks. Implicitly, I assume MFs can borrow overnight at $t = 1$ at the risk-free rate, which is normalized to equal one. They then fully repay any overnight credit lines at $t = 1^+$ by liquidating portfolio assets. There is no default. The existence of these credit lines is only a technical feature of the model resulting from the assumption that ETF and MF trades settle instantaneously but funds do not hold cash: it has no implications on the model predictions. The only purpose of bank credit lines is to bridge the gap between the time at which MFs must satisfy net redemptions by investors in cash, at $t = 1$, and when trading in index markets starts again at the beginning of the subsequent period, $t = 1^+$. Eventually, MFs must eventually liquidate portfolio securities to meet net fund redemptions.

Over time, the total number of fund shares outstanding changes as investors redeem shares. Since investors cannot access the risk-free asset at $t = 0$, the shares outstanding for ETFs and MFs are decreasing over time. Impatient investors are forced to redeem their fund shares at $t = 1$. Meanwhile, patient investors face leverage constraints, so they cannot acquire additional shares beyond their initial investment. Their options are limited to redeeming shares early or remaining invested until the terminal period. Consequently, the size of both ETF and MF markets shrinks as the economy nears its terminal period.

2.4 Model timeline

The sequence of events and actions in the model is as follows. At $t = 0$, each investor i is born with an endowment of one unit of capital, $\theta_0^i = 1$. Each investor observes her idiosyncratic

[11] While Giuzio, Grill, Kryczka, and Weistroffer (2021) argue that liquid asset reserves can limit run risks and costly liquidations of illiquid asset holdings, Zeng (2017) shows that the predictable re-building of cash reserves ex-post outflows may further exacerbates run risks. Empirically, Jiang and Wang (2021) find that corporate bond funds are reluctant to liquidate their most liquid asset holdings during period of market stress.

liquidity risk, λ_i, as well as the unconditional distribution of the fundamental index value $x_j \sim N(\mu_j, \sigma_j^2)$. Investors then pool their endowments to form ETFs and open-end MFs, in exchange for fund shares. Fund shares are equity claims on funds' assets. Neither MFs nor ETFs hold any cash, $C_0^{E,j} = C_0^{M,j} = 0$; all shareholder capital is fully invested in the index.

In the interim period, $t = 1$, the fundamental x_j is realized and observed by everyone. At the same time, investors privately learn about the realization of their liquidity shocks. All investors who receive a liquidity shock liquidate their entire fund portfolios immediately. Remaining, patient investors retain the flexibility to either partially liquidate their fund holdings prematurely and invest the proceeds in the risk-free asset until the terminal period, or maintain unchanged fund portfolios until $t = 2$. Investors submit orders in MF and ETF markets, taking as given the MF NAV, $P_1^{M,j}$, and ETF price $P_1^{E,j}$.

In response to the net supply of ETF shares from investors, the AP determines the quantity of ETF shares to redeem and initiates corresponding offsetting orders to index market makers for the composite security. APs trade in financial markets intraday at $t = 1$, concurrently with ETF investors. In contrast, the MF conducts index trades at $t = 1^+$ to fulfill its obligations to redeeming shareholders by $t = 1$ after observing their aggregate net redemptions.

Over the long term, at $t = 2$, the index pays a terminal dividend, denoted as $P_2^j = x_j$. The final payoff for investors is contingent upon their proceeds from premature fund liquidations at $t = 1$ as well as their residual share holdings in ETFs and MFs and the terminal portfolio value of the investment funds. The model sub-periods do not have uniform durations. Instead, the time span between $t = 0$ and $t = 1$, which I refer to as the short term, should be interpreted as a horizon of days or weeks. The intervals between $t = 1$, more precisely $t = 1^+$, and $t = 2$ should be regarded as the long-term period, comprising months or years.

3 Optimal portfolio allocation between ETFs and MFs

Solving the model entails finding investors' optimal portfolio allocations to ETFs and MFs as a function of their liquidity risk, λ^i, and deriving the equilibrium sizes of the ETF and MF sectors across market segments j. The two-period equilibrium is a perfect Bayesian equilibrium. In every period, $t = 0$ and $t = 1$, conditional on her endowment and (expected) liquidity needs, each investor i chooses her optimal allocation to ETFs and MFs, $\theta_t^{i,E,j}$ and $\theta_t^{i,M,j}$, to maximize her expected utility from terminal wealth, taking as given fund prices as well as other investors' portfolio strategies. I focus on pure strategy equilibria because mixed strategy equilibria, in which investors randomly choose their allocation between ETFs and MFs with some probabilities, are not economically meaningful in this context. Formally:

Definition 2 *Within any index segment, j, given $\mu_j, \sigma_j, c_j, \phi_j, \psi, \eta$ and λ_i, a two-stage pure-*

strategy perfect Bayesian equilibrium (PBE) is defined as a sequence of portfolio allocations $\{\theta_0^{i,E,j}, \theta_0^{i,M,j}, \theta_1^{i,E,j}(x_j), \theta_1^{i,M,j}(x_j)\} \in [0,1]$ *and asset prices,* $P_1^j, P_{1+}^j, P_1^{E,j}, P_2^{E,j}, P_1^{M,j}, P_2^{M,j}$*, such that*

(i) given other investors' allocations at $t = 1$ *and the conditional probability of a mutual fund run,* $\pi^{Run}(x_j)$*, investor* i*'s portfolio allocation strategy* $\{\theta_1^{i,E,j}(x_j), \theta_1^{i,M,j}(x_j)\}$ *maximizes her expected utility 3 at* $t = 1$*, and*

(ii) given investor i*'s expectation of other investors' sequential allocation strategies, including the probability of a mutual fund run at* $t = 1$*,* i*'s allocation* $\{\theta_0^{E,i,j}, \theta_0^{M,i,j}\}$ *maximizes her expected utility 3 at* $t = 0$*, and*

(iii) all investors share common beliefs about the probability of a mutual fund run at $t = 1$ *conditional on the realization of* x_j*, and*

(iv) ETF and index markets clear at all times.

I solve for the equilibrium of the model using backward induction. First, I take as given the realization of the fundamental index value x_j at $t = 1$, as well as investors' initial allocations, $\theta_0^{E,i,j}$ and $\theta_0^{M,i,j}$, and I then solve for the MF, ETF, and index market equilibrium at $t = 1$. In the baseline model in which APs and MFs trade with separate index market makers, the ETF and MF market equilibria at $t = 1$ are not interdependent and can be solved separately. This follows from the assumption that investors can no longer switch between fund types in the interim period. MF prices are fixed at the fund NAV over the short term and do not instantaneously adjust to demand-supply imbalances in MF markets. Hence, MF investors takes prices as given. They do not internalize the fund-level transaction costs caused by their trades. In contrast, ETF investors take into account their aggregate price impact in ETF markets as they trade intraday at the fully flexible market-clearing ETF price.

A complication emerges because patient MF investors' optimal allocation at $t = 1$ depends on their beliefs about other investors' redemption decisions. The interdependency of investors' allocation decisions and the potential for shareholder runs on MFs results in multiple equilibria in MF markets. I use the sunspot equilibrium selection technique to coordinate MF investor behavior and select the unique MF market equilibrium at $t = 1$. The notion of sunspot equilibria originated with Cass and Shell (1983) and has found contemporary applications in financial intermediation models, including recent work by Dávila and Goldstein (2023). Specifically, within the realm of index fundamentals, x_j, where multiple equilibria can arise in MF markets, I assume an equal probability for both the run and no-run equilibria. Within this framework, the run-equilibrium entails early redemption of all patient MF investors' shares at $t = 1$, while the no-run equilibrium involves patient MF investors remaining invested until the terminal period.

An alternative method to address the challenge of multiple equilibria is the global games technique introduced by Morris and Shin (2003) and further explored by Goldstein and Pauzner

(2005) for panic-based bank runs. This approach has previously been employed to identify a unique equilibrium in context of mutual fund runs (Chen et al. 2010) but necessitates asymmetric information about fundamentals among investors, which would compromise the model's tractability. Since this paper's primary focus does not lie on the MF equilibrium solution, I opt for the more pragmatic sunspot equilibrium selection technique.

Second, taking as given investors' optimal investment policy and equilibrium asset prices as a function of x_j at $t = 1$, I solve for investor equilibrium allocations at $t = 0$ as a function of their idiosyncratic liquidity risks, λ^i.

3.1 Fund payoffs in the terminal period

Asset prices at $t = 2$ directly follow from the terminal index payoff and are given by:

$$P_2^j = x_j, \tag{13}$$

$$P_2^{E,j} = P_2^j, \tag{14}$$

$$P_2^{M,j} = \frac{X_2^{M,j} P_2^j}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}. \tag{15}$$

The terminal ETF price in equation 14 equals the fundamental value of its benchmark index given by equation 13. Over the long term, APs face no balance sheet capacity constraints, and they resume ETF arbitrage until all relative mispricing has been eliminated. This is akin to asserting that, even in less liquid markets, the law of one price must eventually prevail.

By design, the terminal MF payoff in equation 15 is equal to the fund NAV at $t = 2$. The fund NAV is defined as the value of the fund's portfolio assets divided by its shares outstanding. $\kappa_0^{M,j}$ is the number of MF shares outstanding at $t = 0$, which is equivalent to the proportion of investors who initially invest in MFs, and $\Delta\kappa_1^{M,j} = \kappa_0^M - \kappa_1^M$ is the number of MF shares redeemed in the interim period, so $\kappa_1^M = \kappa_0^M - \Delta\kappa_1^{M,j}$. If the MF experiences net outflows: $\Delta\kappa_1^{M,j} > 0$. $0 \leq X_2^{M,j} \leq \kappa_0^M$ is the number of index shares held by the MF at $t = 2$ after accounting for the fund's interim index trades. All variables on the right-hand side of equation 15 are known as of $t = 1^+$. Since one MF share initially represents a claim on one unit of the index, the long-term MF payoff deviates from the terminal index value when $X_2^{M,j} \neq \kappa_1^M$. I refer to these deviations between the fund and index payoff as tracking difference.

Definition 3 *The MF tracking difference denotes the difference between the mutual fund net asset value (NAV) and the benchmark index price at any time and is given by:*

$$\Delta_t^{M,j} \equiv P_t^j - NAV_t^{M,j}. \tag{16}$$

When $\Delta_t^{M,j} > 0$, the MF NAV is smaller than the value of its benchmark index: The MF exhibits a positive tracking difference.

Δ_t^M represents a wedge between the fund and index price. As evident from the definition of $NAV_t^{M,j}$ in equation 15, $\Delta_t^M \neq 0$ arises from the cumulative transaction costs incurred at the fund level as a result of early redemptions by MF shareholders. Equation 16 defines the tracking difference as a difference in prices. In practice, the tracking difference typically pertains to deviations of fund returns from a benchmark return. Because all asset prices are normalized to equal one in the initial period, prices and returns are equivalent in the model.

Following lemma 1, for ETFs, the tracking difference is always zero, $\Delta_t^{E,j} = 0 \ \forall t$. ETF investors trading in secondary ETF markets bear their own transaction costs. Their actions impact the ETF's market price but do not affect the ETF's net asset value. Outflows from ETFs only have temporary effects on investors' payoffs, whereas outflows from MFs can have persistent effects on investors' payoffs.

Lemma 2 *In the terminal period the MF tracking difference per MF share is given by:*

$$\Delta_2^{M,j} = \frac{P_2^j \left(\frac{\Delta\kappa_1^{M,j} (NAV_1^{M,j} - P_{1^+}^j)}{P_{1^+}^j} \right)}{\kappa_1^{M,j}}. \tag{17}$$

The relative MF tracking difference, $\tilde{\Delta}_2^{M,j} = \frac{\Delta_2^{M,j}}{P_2^j}$, per MF share is given by

$$\tilde{\Delta}_2^{M,j} = \underbrace{\frac{1}{\kappa_1^{M,j}}}_{\substack{\text{Remaining} \\ \text{MF shares} \\ \text{outstanding}}} \left(\underbrace{\Delta\kappa_1^{M,j}}_{\substack{\text{\# MF} \\ \text{shares} \\ \text{redeemed} \\ \text{early}}} \left(\underbrace{NAV_1^{M,j} (P_{1^+}^{j,M})^{-1} - 1}_{\substack{\text{Excess \# index} \\ \text{shares liquidated} \\ \text{to satisfy early} \\ \text{redemptions}}} \right) \right).$$

Corollary 1 *The MF tracking difference in the terminal period is zero, $\Delta_2^{M,j} = 0$, if and only if at least one of the following three conditions is satisfied:*

(i) Markets are perfectly liquid and funds have no price impact when trading index shares, $c_j = 0$ and $\psi = 1$.

(ii) The mutual fund NAV at $t = 1$ is perfectly forward looking and equal to the index price at which the MF trades in index markets, $NAV_1^{M,j} = P_{1^+}^{M,j}$.

(iii) Net mutual fund redemptions at $t = 1$ are zero, $\Delta\kappa_1^{M,j} = 0$.

If condition (i) is satisfied and MFs have no price impact in index markets, (ii) automatically follows. In this case $NAV_1^{M,j} = P_{1^+}^j = x_j$. Early investors trade the MF at a price reflecting its fair value after accounting for (zero) price impact. They do not impose any negative externalities on the remaining MF investors. Condition (ii) may for example hold if the MF used

swing pricing and swing factors were calibrated optimally to reflect all flow-induced transaction costs. I study the implications of swing pricing on MF payoffs and investors' allocation decision in section 4. Whereas (i) implies (ii), the reverse is not true. Finally, condition (iii) illustrates that fund flows are necessary for the costs of the liquidity-transformation services provided by investment funds to materialize.

Corollary 2 *When at least one of the conditions in corollary 1 is satisfied, the mutual fund NAV at $t = 2$ is equal to the terminal index payoff. In this case, the MF and ETF tracking benchmark index j provide identical payoffs over the long term, $P_2^{M,j} = P_2^{E,j} = P_2^j$.*

Equation 17 illustrates how frictions in MFs arise from flow-induced share dilution. The imperfect flexibility of the MF NAV relative to prices in underlying security markets, is the fundamental source of tracking difference, $\Delta_2^{M,j} \neq 0$, over the long term. When $NAV_1^{M,j} - P_{1^+}^j > 0$, short-term investors' payoffs exceed the current liquidation value of their share holdings. They get a better deal than they would if they had to bear the full cost associated with their trading activity. Accordingly, the MF provides short-term investors with liquidity insurance in the amount of $NAV_1^{M,j} - P_{1^+}^j$. This liquidity provision to short-term investors is not free since the relatively higher payoff received by short-term investors comes at the cost of long-term investors. In this sense, $\Delta_2^{M,j}$ can also be interpreted as a negative risk premium or an "insurance" premium paid in exchange for the short-term liquidity-insurance service provided by MFs. It represents a redistribution of consumption from long-term to short-term investors.

3.2 Portfolio allocations in the interim period

At $t = 1$, the idiosyncratic liquidity shocks are realized. Each agent learns if she is the impatient (e) or patient (l) type. Besides, the fundamental index value x_j is revealed.

Lemma 3 *Conditional on the realization of the liquidity shocks at $t = 1$, there are four groups of ex-post identical investors: patient and impatient ETF investors, and patient and impatient MF investors. These investors are characterized by the following conditions:*

(i) Patient ETF investors: All $i = l$ who initially invested in ETFs, $\theta_0^{i,E,j} = 1$

(ii) Impatient ETF investors: All $i = e$ who initially invested in ETFs, $\theta_0^{i,E,j} = 1$.

(iii) Patient MF investors: All $i = l$ who initially invested in MFs, $\theta_0^{i,M,j} = 1$.

(iv) Impatient MF investors: All $i = e$ who initially invested in MFs, $\theta_0^{i,M,j} = 1$.

Solving for the fund market equilibrium at $t = 1$ entails separately deriving the optimal portfolio choice of each of these investor groups.

(i) **Patient ETF investors** must decide if they wish to liquidate any of their ETF shares early. Formally, their problem is given by:

$$\max_{\theta_1^{i,E,j}} E[w_2^i|x] \tag{18}$$
$$s.t.\ w_2^i = \theta_1^{i,E,j} P_2^{E,j} + (\theta_0^{i,E,j} - \theta_1^{i,E,j}) P_1^{E,j} R^f,$$
$$\theta_0^{i,E,j} - \theta_1^{i,E,j} \geq 0,$$
$$\theta_1^{i,E,j} \geq 0.$$

The optimization problem 18 follows directly from the assumption that investors are no longer able to switch between fund types in the interim period. Rational investors anticipate that the ETF mispricing will converge to zero in the terminal period, and $P_2^{E,j} = x_j$. There remains no uncertainty regarding the final ETF payoff. Hence, patient ETF investors' optimal portfolio allocation at $t = 1$ depends only on the current ETF price, index value, and R^f.

Assumption 3 *If the terminal ETF payoff is equal to the payoff from immediate liquidation at $t = 1$, patient investors always hold the investment fund until maturity.*

Using assumption 3, patient ETF investors' optimal investment policy is given by:

$$\theta_1^{i,E,j} = \begin{cases} 0 & \text{if } P_2^j - (P_1^j - \epsilon_1^{E,j}) R^f < 0, \\ 1 & \text{if } P_2^j - (P_1^j - \epsilon_1^{E,j}) R^f \geq 0, \end{cases}$$

The ETF discount $\epsilon_1^{E,j}$ is the liquidity premium that ETF investors pay in exchange for short-term liquidity provision. The larger the current ETF discount, the greater the payoff from waiting to liquidate ETF shares until the terminal period. If the expected payoff of the ETF in the terminal period is smaller than the current ETF share price reinvested at the risk-free rate, $x_j < P_1^E R^f$, investors always liquidate all of their ETF shares prematurely. In the absence of ETF inflows, this condition is never satisfied in equilibrium.

Proposition 1 *In equilibrium, patient investors never liquidate any ETF shares early:*

$$\theta_1^{i,E,j*} = 1 \ \forall i = l \ \textit{with} \ \theta_0^{i,E,j} = 1.$$

(ii) **Impatient ETF investors'** terminal wealth is given by:

$$w_1^i = P_1^{E,j}.$$

By definition, $u(c_2^i) = u(w_1^i)\ \forall i = e$.

(iii) **Patient MF investors** can decide to liquidate any of their MF shares early like patient ETF investors. They solve the following problem:

$$\max_{\theta_1^{i,M,j}} E[w_2^i|x] \tag{19}$$
$$s.t.\ w_2^i = \theta_1^{i,M,j} P_2^{M,j} + (\theta_0^{i,M,j} - \theta_1^{i,M,j}) P_1^{M,j} R^f,$$
$$\theta_0^{i,M,j} - \theta_1^{i,M,j} \geq 0,$$
$$\theta_1^{i,M,j} \geq 0.$$

In contrast to the terminal ETF payoff, the terminal MF payoff is uncertain even at $t = 1$. Beyond x_j, $P_2^{M,j}$ depends on the early redemption decision of other MF investors. This interdependency arises because the MF tracking difference, $\Delta_2^{M,j}$, depends on fund flows at $t = 1$. By definition, $E[P_2^{M,j}|x_j] = x_j - E[\Delta_2^{M,j}|x_j]$. Because investors are risk-neutral, the solution to equation 19 will always be on the boundary of the permissible range of $\theta_1^{i,M,j}$.

Patient MF investors' optimal allocation policy at $t = 1$ as a function of the expected MF tracking difference is then given by:

$$\theta_1^{i,M,j} = \begin{cases} 0 & \text{if } P_2^j - E[\Delta_2^{M,j}|x] - P_1^{M,j} R^f \leq 0, \\ 1 & \text{if } P_2^j - E[\Delta_2^{M,j}|x] - P_1^{M,j} R^f \geq 0. \end{cases} \tag{20}$$

In equilibrium, investors are perfectly forward looking and anticipate other investors' actions conditional on the realization x_j. In the region of x_j characterized by the potential existence of multiple mutual fund market equilibria, the anchoring of expectations is achieved through a sunspot equilibrium.

(iv) **Impatient MF investors'** final wealth is given by:

$$w_1^i = \psi x_j.$$

3.3 Equilibrium prices in the interim period

Equilibrium prices in the interim period are determined by market clearing between the MF, AP, index market makers and ETF investors, given MF investors' fund liquidations. In the baseline model with segmented index markets, prices are set in the following sequence: First, at $t = 1$ the ETF price, P_1^{E,j^*}, is determined relative to the index price from the AP's optimal ETF redemptions and investors' net supply of ETF shares. Simultaneously, the index price faced by the AP, $P_1^{j^*}$, follows from the other side of its arbitrage trade and the market maker's net demand for index shares. Second, given MF investors' net redemptions at $t = 1$, at $t = 1^+$ the index price faced by MFs, $P_{1^+}^{j\ *}$ is determined from market clearing between the MF's flow-induced index sales and its market maker's net demand for index shares. Thereby, $P_{1^+}^{j\ *}$ is derived independently of $P_1^{j^*}$ due to the timing disparities between the intraday trading activity of the AP and the next-day index trading activity of the MF. The model extension

with integrated index markets is presented in section 3.3.1.

ETF price. $P_1^{E,j}$ follows from market clearing between APs and ETF investors, taking as given the equilibrium index price P_1^j. Since patient ETF investors never redeem early, the aggregate number of ETF shares sold by investors at $t = 1$, denoted by $\Delta\kappa_1^{E,j}$, is equal to impatient investors' liquidations:

$$\Delta\kappa_1^{E,j} = \int_i (\theta_0^{i,E,j^*} - \theta_1^{i,E,j^*})\ di.$$

From equation 10 follows the AP's net demand function for ETF shares at $t = 1$:

$$\Theta_1^{E,j,AP} = \frac{(P_1^j - P_1^{E,j})}{\phi_j}. \tag{21}$$

$\Theta_1^{E,j,AP}$ refers to the number of ETF shares redeemed by APs in exchange for an equivalent amount of index shares. Equation 21 implies that when the ETF trades at a discount (premium), the AP opts to redeem (create) ETF shares to the extent permitted by its balance sheet capacity. ETF market clearing requires $\Theta_1^{E,j,AP} = \Delta\kappa_1^{E,j}$. Hence, given the index price and ETF outflows by impatient investors, the ETF price as a function of the index price is:

$$P_1^{E,j} = P_1^j - \Delta\kappa_1^{E,j}\phi_j.$$

Index price at $t = 1$. Conditional on market clearing in ETF markets, the index price faced by the AP at $t = 1$ follows from market clearing with the index market maker. $\Theta_1^{AP,j} = \Theta_1^{E,AP}$ since the AP can always redeem one unit of the index in exchange for one ETF share. In the case of $c_j > 0$, when there is price impact in index markets, the $t = 1$ index price is:

$$P_1^j = E[P_2^j|x_j] - \Delta\kappa_1^{E,j}c_j.$$

For the ETF price it implies:

$$P_1^{E,j} = E[P_2^j|x_j] - \Delta\kappa_1^{E,j}(c_j + \phi_j). \tag{22}$$

ETF prices adjust in real time to supply and demand conditions in financial markets. ETF investors internalize the price impact in index markets caused by AP arbitrage as well as APs' balance sheet capacity constraints. Because of AP balance sheet capacity constraints, $\phi_j > 0$, ETF prices are excessively flexible over the short term, $P_1^{E,j} < P_1^j$.

In the alternative scenario with $c_j = 0$, the market maker's net demand for index shares is perfectly elastic. Index markets are infinitely liquid. APs do not have price impact. This condition represents the reference point for liquid market segments, such as the S&P 500 index, during periods of abundant market-wide liquidity. In this setting, both APs and MFs face a common equilibrium index price at $t = 1$ and $t = 1^+$, represented by $P_1^j = P_{1^+}^j = x_j$.

Proposition 2 *The payoffs of the ETF and MF tracking index j are identical across all investment horizons, $P_t^{E,j} = P_t^{M,j}$ $\forall t = 1, 2$ if one of the following two conditions is satisfied:*

(i) Index markets are perfectly liquid and APs do not face any balance sheet capacity constraints, $c_j = 0$, $\psi = 0$ and $\phi_j = 0$.

(ii) Index markets are illiquid ($c_j > 0$), but the MF NAV at $t = 1$ is perfectly forward looking, APs do not face any balance sheet capacity constraints and the volume of outflows from ETFs and MFs at $t = 1$ are equal, $\Delta\kappa_1^{E,j} = \Delta\kappa_1^{M,j}$.

In these idealized environments, a fund structure irrelevance principle emerges: Under assumptions (i) or (ii), the capital structure of a passive investment fund – if it is structured as an ETF or an open-end mutual fund, has no impact on the payoffs to its investors.

Proposition 2 constitutes a special case of my theory. On one side, the assumptions that support condition *(i)* may be applicable to large-cap domestic equity index funds in times of ample market liquidity. On the other side, the requirement of equivalent volumes of fund outflows from the ETF and MF *(ii)* is unlikely satisfied due to the random nature of investors' liquidity shocks. There are two possible modifications of *(ii)*: First, even if the ETF and MF tracking index j experience different outflows ex-post at $t = 1$, their ex-ante expected payoffs are identical across investment horizons as long as investors self-select into both fund types at $t = 0$ at random, irrespective of their λ_i. Second, if the ETF and MF trade simultaneously in the index markets during the interim period, they encounter identical index prices and offer equivalent payoffs to their investors, regardless of their individual outflows. The latter result necessitates that MFs accurately predict their investors' net redemption requests.

Mutual fund price. Under the baseline specification, the MF NAV at $t = 1$, is given by:

$$P_1^{M,j} \equiv \psi E[P_2^j | x_j], \tag{23}$$

, where $0 < \psi < 1$ reflects a penalty for premature fund liquidations. Formally, $\psi < 1$ ensures that, in the absence of payoff complementarities, early redemptions would never be optimal for patient investors. Investors earn a positive expected return when holding fund shares until the underlying asset matures in $t = 2$.

In practice, trades of most U.S. MFs, with the exception of money market mutual funds (MMFs), settle within one business day after the trade date, $T + 1$, whereas ETF transactions currently settle $T + 2$.[12] Therefore, in the case of MF redemptions, investors receive the cash proceeds from their sales with a delay of at most one business day. In my model, investors

[12] See *SEC settlement cycle recommendation*. Generally, the settlement cycle for transactions of publicly traded securities, including ETFs, in the U.S. is $T + 2$. MMFs tend to settle $T + 0$ or $T + 1$. On February 15, 2023, the SEC adopted an amendment to an existing rule to further reduce the settlement cycle for standard securities transactions to $T + 1$, see *SEC release Nos. 34-96930, IA-6239; File No. S7-05-22*.

receive the cash from their ETF and MF transactions on the same day, a convention followed for simplicity. This assumption does not affect the model predictions; rather, accounting for the later settlement date of ETF compared to MF transactions strengthens model predictions.

The number of MF redemptions are $\Delta\kappa_1^{M,j} = \int_i(\theta_0^{i,M,j} - \theta_1^{i,M,j})di$. It is noteworthy that even in the case of a MF run in which all patient MF investors redeem their shares early, the presence of the "sticky" MF investors ensures that the MF never fully disappears at $t = 1$, $\kappa_t^M \geq \eta > 0$. This assumption does not mechanically produce the model predictions. It is only necessary to simplify the model solution as it allows to abstract from events in which the MF goes bankrupt. The equilibrium coexistence of ETFs and MFs follows from investors initial allocation decisions, $\theta_{i,0}^M$ and $\theta_{i,0}^E$, which are fully endogenous in the model.

Index price at $t = 1^+$. The index price at $t = 1^+$ follows from market clearing between the market maker's net demand for index shares, $\Theta_{1^+}^{D,j} = \frac{E[P_2^j|x_j] - P_{1^+}^j}{c_j}$, and MFs flow-induced trading in index markets:

$$\Theta_{1^+}^{M,j} = \frac{\Delta\kappa_1^{M,j} P_1^{M,j}}{P_{1^+}^j}. \tag{24}$$

The maximum number of index shares which the MF can sell to meet investor redemptions is bounded by its portfolio holdings according to $\Theta_{1^+}^{M,j} \leq \kappa_0^M$ due to short sale constraints.

In the special case in which there is no price impact in index markets, $c_j = 0$, the number of index shares sold by MFs is given by $\Theta_{1^+}^{M,j} = \psi\Delta\kappa_1^{M,j}$. Abstracting from potential taxable distributions of capital gains and trading commissions, there are no transaction costs and therefore externalities associated with fund outflows.

Lemma 4 *If index markets are frictionless, $c_j = 0$, patient MF investors never choose to redeem any MF shares early in equilibrium. There are no run risks in the index MF.*

This result is consistent with prior empirical studies documenting higher run risk in more illiquid fund market segments (e.g., Chen et al. (2010)).

In the more general case in which $0 < c_j \leq 1$, investment funds have price impact in index markets. The market clearing index price at $t = 1^+$, follows from $\Theta_{1^+}^{M,j} = \Theta_{1^+}^{D,j}$ and solves:

$$\frac{E[P_2^j|x] - P_{1^+}^j}{c_j} = \frac{\Delta\kappa_1^{M,j} P_1^{M,j}}{P_{1^+}^j}. \tag{25}$$

It exists as long as the fundamental index value x is large relative to the market maker's inventory costs, c_j. To establish the uniqueness, I impose an additional assumption.

Assumption 4 *If there exists a pair of candidate prices and quantities,* $\{P'_{1^+}, \Theta'_{1^+}\}\{P''_{1^+}, \Theta''_{1^+}\}$*, satisfying equation 25, the equilibrium index price is given by the larger price candidate:*

$$P^j_{1^+} = P'_1 \text{ if } P'_1 \geq P''_1,$$

and $\Theta^{M,j}_{1^+} = \Theta'_1$.

Assumption 4 is consistent with a model in which market makers and MF managers engage in price negotiations for their trade, defined by $\{P'_1, \Theta'_1\}$. They start at $P^j = x_j$ and adjust the index price downward until the market clears.

Proposition 3 *If* $c_j > 0$ *and* $\frac{x_j - \frac{1}{2}(x_j - \sqrt{x_j^2 - 4c_j\psi x_j \Delta\kappa_1^{M,j}})}{c_j} \leq 1 + \eta$*, for any volume of MF redemptions* $\Delta\kappa_1^{M,j} \in [0,1]$*, the unique index market equilibrium at* $t = 1^+$ *is given by:*

$$P^j_{1^+} = \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j\psi x_j \Delta\kappa_1^{M,j}}),$$
$$\Theta^j_{1^+} = \frac{x_j + \frac{1}{2}(x_j - \sqrt{x_j^2 - 4c_j\psi x_j \Delta\kappa_1^{M,j}})}{c_j}.$$

The condition $\frac{x_j - \frac{1}{2}(x_j - \sqrt{x_j^2 - 4c_j\psi x_j \Delta\kappa_1^{M,j}})}{c_j} \leq 1 + \eta$ follows from $\Theta^{M,j}_{1^+} \leq \kappa_0^{M,j}$.

Corollary 3 *In the case with* $c_j > 0$*, there exists an market clearing index price* $P^j_{1^+}$ *that solves 25 over* $0 \leq \Theta^{M,j}_{1^+} \leq 1 + \eta$ *if and only if* $x_j \geq 4c_j\psi$.

Corollary 4 *In the special case in which MF net redemptions at* $t = 1$ *are given by:*

$$\Delta\kappa_1^{M,j} = \frac{(1-\psi)x_j}{c_j},$$

the equilibrium index price at $t = 1^+$ *is exactly equal to the MF net asset value at* $t = 1$*,* $P^j_{1^+} = \psi x_j$*. Then,* $\Theta^j_{1^+} = \Delta\kappa_1^{M,j}$*, and the MF tracking difference is zero,* $\Delta_2^{M,j} = 0$*. In this special case, the fund NAV at* $t = 1$ *is perfectly forward looking.*

Ex-post, this equilibrium is comparable to the case in which the MF employs swing pricing to ensure that exiting investors at $t = 1$ bear the transaction costs associated with their redemptions. Ex-ante it differs from a swing pricing equilibrium due to the effect of swing pricing policies on investors' expectations regarding others' redemption decisions.

Corollary 5 *When* $\Delta\kappa_1^{M,j} \neq \frac{(1-\psi)x_j}{c_j}$*, the equilibrium index price at which the MF trades at* $t = 1^+$ *deviates from the MF NAV at which fund investors can redeem shares at* $t = 1$*:*

i. If $\Delta\kappa_1^{M,j} < \frac{(1-\psi)x_j}{c_j}$*, the equilibrium index price at* $t = 1^+$ *is larger than the MF NAV at* $t = 1$*,* $P_{1^+}^j > \psi x_j$*. Then,* $\Theta_{1^+}^{M,j} < \Delta\kappa_1^{M,j}$*, and the terminal MF tracking difference becomes negative,* $\Delta_2^{M,j} < 0$*. The remaining MF shareholders gain from other investors' redemptions at* $P_1^{M,j}$ *because the latter sell at a fund NAV that is too low given fundamentals. Premature redemptions are costly.*

ii. If $\Delta\kappa_1^{M,j} > \frac{(1-\psi)x_j}{c_j}$*, the equilibrium index price at* $t = 1^+$ *is lower than the MF NAV at* $t = 1$*,* $P^j{}_{1^+} < \psi x$*. Then,* $\Theta_{1^+}^{M,j} > \Delta\kappa_1^{M,j}$*, and the terminal MF tracking difference becomes positive,* $\Delta_2^{M,j} > 0$*. The MF NAV at* $t = 1$ *is too high, as it does not fully account for the full price impact associated with investor redemptions at* $t = 1$*.*

3.3.1 Fund and index prices with a common index market maker

In an alternative specification, index markets are integrated and APs and MFs sequentially trade with the same index market maker. In this case funds compete on the liability as well as asset side. Because ETF arbitrage takes place intraday, the AP trades first at $t = 1$. Next, given the price impact generated by the AP's arbitrage activities, the MF trades with the index market maker at $t = 1^+$. The net demand schedule for index shares faced by the AP remains the same as before. As a result, the index and ETF price at $t = 1$ are the same under sequential trading with a common index market maker as in the baseline specification with segmented index markets:

$$P_1^{j,Seq} = x_j - \Delta\kappa_1^{E,j} c_j = P_1^j,$$
$$P_1^{E,j,Seq} = x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j) = P_1^{E,j}.$$

The MF price at $t = 1$ as well as the index price faced by the MF at $t = 1^+$ are different with sequential index trading: First, the MF NAV is set at the end of trading day, so the AP's intraday trading activities directly impact the MF price paid to redeeming investors at $t = 1$. Whereas the MF NAV was quasi-exogenous in the baseline specification, it is now fully endogenous and equal to the marginal index price faced by the AP:

$$P_1^{M,j,Seq} = x_j - \Delta\kappa_1^{E,j} c_j.$$

As long as $x_j(1-\psi) < c_j \Delta\kappa_1^{E,j}$, that is c_j is large and ψ close to one, it holds that:

$$P_1^{M,j,Seq} < P_1^{M,j}$$

ETF outflows do not only negatively impact ETF prices, through the AP arbitrage channel, they also decrease the end-of-day NAVs of same-index MFs to the extent that APs pass on selling pressure from ETF to index markets. The intraday nature of ETF trading makes MF prices endogenously more flexible. Redeeming MF investors at $t = 1$ still do not face the

trading costs associated with their own redemptions, but they internalize the price impact caused by impatient ETF investors' early liquidations.

Second, the net demand schedule for index shares faced by the MF differs:

$$\Theta_{1^+}^{D,j,Seq} = \frac{x_j - \Delta\kappa_1^{E,j} c_j - P_{1^+}^j}{c_j}.$$

The market maker's index inventory expands after trading with the AP. Given its quadratic inventory cost, the market maker therefore demands an even larger discount relative to the fundamental index value, x_j, to absorb the additional supply of index shares from the MF.

Overall, asset side competition between same-index ETFs and MFs reduces the MF's short-term liquidity provision, $P_1^{M,j,Seq} < P_1^{M,j}$, to the benefit of long-term MF investors. The greater flexibility of the MF price at $t = 1$ discourages early redemptions by patient MF investors as they are forced to bear the transient price impact cause by ETFs in index markets. ETF investors' trading at $t = 1$ comes at the cost of short-term MF investors to the benefit of long-term MF investors.

The key model predictions regarding investors' optimal allocation and funds relative liquidity provision over time continue to hold. For tractability, the rest of the paper builds on the simplified baseline model with segmented index markets. This is consistent with the practice of broker-dealers to have separate trading desks responsible for ETF arbitrage as compared to market making for mutual funds and other buy-side entities.

3.4 Equilibrium allocations in the interim period

Investors' equilibrium portfolio allocations in the interim period, θ_1^{i,E,j^*} and θ_1^{i,M,j^*}, directly follow from market clearing in the ETF, MF, and index markets at $t = 1$ and investors' optimal investment strategies.

ETF market equilibrium. From proposition 1, it follows:

$$\theta_1^{i,E,j^*} = \begin{cases} 1 & \forall i = l \text{ (patient types) with } \theta_0^{i,E,j^*} = 1, \\ 0 & \forall i = e \text{ (impatient types) with } \theta_0^{i,E,j^*} = 1 \end{cases}$$

and

$$P_1^{E,j^*} = x_j - \underbrace{\Delta\kappa_1^{E,j^*} c_j}_{\substack{\text{Price} \\ \text{discovery}}} + \underbrace{\Delta\kappa_1^{E,j^*} \phi_j}_{\substack{\text{Relative} \\ \text{mispricing}}}. \tag{26}$$

$\Delta\kappa_1^{E,j^*} = e^E$ is the fraction of ex-post impatient ETF investors.

Besides the fundamental, x_j, the equilibrium ETF price in equation 26 depends on two components, one reflecting price discovery and another relative mispricing. First, the price discovery component is given by $\Delta\kappa_1^{E,j^*} c_j$. It reflects the price impact generated in index markets as a result of the AP's arbitrage trades. When deciding on its optimal ETF redemptions, the AP takes into account its expected price impact in index markets and directly passes it on to the trading ETF investors. Through this mechanism, price discovery takes place in ETF markets intraday. Early ETF investors pay the flow-induced transaction costs caused by their trading decisions.

Second, the equilibrium relative ETF mispricing is given by

$$\epsilon^{E,j^*} = \Delta\kappa_1^{E,j^*} \phi_j.$$

Since $e^E \geq 0$ whenever the ETF has non-zero AUM and $\phi_j > 0 \ \forall j$, the ETF trades at a discount to its benchmark index. The equilibrium mispricing is increasing in $\Delta\kappa_1^{E,j^*}$ and ϕ_j. It arises because ETF arbitrage temporarily consumes the AP's balance sheet capacity. The more illiquid the ETF's index holdings (higher ϕ_j), the more time it takes the AP to complete an ETF arbitrage trade and therefore, the longer the AP has to hold the index shares on its balance sheet. The latter aspect of the model mirrors the empirical fact that corporate bond and international equity ETFs often exhibit larger mispricing when compared to large-cap domestic equity ETFs (see figures A.5 - A.12). Intuitively, amid excess supply of ETF shares from investors, the ETF price has to adjust downward to incentivize the AP to step in and supply liquidity on secondary ETF markets. $\epsilon_1^{E,j}$ represents the AP's compensation for liquidity provision in times of aggregate illiquidity. An alternative approach to generating the same results is by making the AP risk-averse and introducing price risk in index markets.

As a result, the ETF price is excessively flexible over the short term. It does not only reflect index market price impact but also the balance sheet cost associated with AP's liquidity-provision services. Both costs are borne by impatient ETF investors. Investors cannot create or redeem ETF shares themselves. They depend on the AP's creation and redemption activities and bear all associated costs. Overall, the more illiquid the index j and the higher the mass of impatient ETF investors at $t = 1$, the larger the index market price impact and ETF mispricing, and thus the liquidation costs for any individual impatient ETF investor.

Patient ETF investors have no incentive to liquidate their shares early. ETF investors only trade fund shares when they need liquidity. In this specification, price impact in index markets is transitory. Over the long term, the ETF pays x_j independent of trading activity in prior periods. Accordingly, by waiting to liquidate until the terminal period, patient ETF investors can avoid both, the index price impact and relative ETF mispricing. The results

are robust to an alternative specification in which index price impact is persistent. ETF mispricing is sufficient to deter early liquidations by patient ETF investors.

MF market equilibrium. Patient MF investors' early redemption incentives illustrated in equation 20 are increasing in the expected volume of redemptions by other MF investors, $E[\Delta\kappa_1^{M,j}|x_j]$, index market illiquidity, c_j, and decreasing in the fundamental, x_j. Each MF investor's optimal allocation is a function of the expected terminal MF tracking difference, which in turn depends on the allocation decisions of all MF investors at $t = 1$. As a result of the payoff complementarities among MF investors, the MF market at $t = 1$ is characterized by the possibility of multiple equilibria. To show this formally, I first analyze two regions of very bad and very good fundamentals, where each patient MF investor's optimal allocation is independent of her beliefs regarding other patient MF investors' actions. Following the previous literature (e.g., Goldstein and Pauzner (2005)), I refer to these regions of fundamental values as the lower and upper dominance regions. In these two regions, patient MF investors' allocations at $t = 1$ are only a function of the fundamental index value and general model parameters.

Lower dominance region. The lower dominance region encompasses values of the fundamental in the range $x_j \in (0, \underline{x}_j]$. In this region, fundamentals are sufficiently bad, such that early redemption is the dominant strategy for any individual patient investor. After observing $x_j < \underline{x}_j$, an investor redeems all her MF shares early, even if all other patient MF investors choose not to redeem their shares early. At the boundary $x_j = \underline{x}_j$, a patient investor is indifferent between redeeming early and remaining invested until $t = 2$. Let $\bar{e}^M = E[e^M]$ be the expected mass of impatient MF investors at $t = 1$ and note that $\bar{e}^M$ only depends on the initial mass of MF investors, $\kappa_0^{M,j}$, and their liquidity risk λ_i.

Then, the lower dominance region is characterized by the value $\underline{x}_j$, which solves:

$$E[\underbrace{P_2^j - \Delta_2^{M,j}}_{\substack{\text{Payoff from}\\ \text{liquidating}\\ \text{at } t=2}} - \underbrace{P_1^{M,j} R^f}_{\substack{\text{Payoff}\\ \text{from early}\\ \text{redemption}}} | x_j = \underline{x}_j \cup \Delta\kappa_1^{M,j} = \bar{e}^M] = 0. \tag{27}$$

All uncertainty in equation 27 comes from $\Delta_2^{M,j}$'s dependence on MF investors' redemptions.

Upper dominance region. The upper dominance region is defined by $x_j \in [\overline{x_j}, \infty)$. In this region, fundamentals are so good that patient MF investors always keep all of their MF shares until $t = 2$. After observing $x_j \geq \overline{x}_j$, they never redeem early even if all other patient MF investors are liquidating at $t = 1$. Before portfolio reallocation, the mass of all patient and impatient MF investors at $t = 1$ is $e^M + l^M \leq 1$. Hence, $e^M + l^M = \kappa_0^{M,j} - \eta$, where η is common knowledge and $\kappa_0^{M,j}$ is observed after allocation decision have been made in the initial period. Then, the upper dominance region is characterized by the value $\overline{x}_j$ that solves:

$$E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \overline{x}_j \cup \Delta\kappa_1^{M,j} = \kappa_0^{M,j} - \eta] = 0 \tag{28}$$

The condition 28 is similar to condition 27 except that it is derived conditional on all other MF investors liquidating their shares early, $\Delta\kappa_1^{M,j} = e^M + l^M$.

Proposition 4 *There exists a lower and upper dominance region with respect to the fundamental that is characterized by the boundary values $\underline{x}_j$ and $\overline{x}_j$, with $\underline{x}_j \leq \overline{x}_j$, such that for any realization of the fundamental within these regions ($x_j \leq \underline{x}_j$ or $x_j \geq \overline{x}_j$) the MF market at $t = 1$ has a unique equilibrium in which patient MF investors always redeem all of their shares early or never redeem any shares early at all, irrespective of their beliefs of other patient MF investors' portfolio allocations. Formally, $\exists\ \underline{x}_j,\ \overline{x}_j$ s.t.*

$$\theta_1^{i,M,j^*}(x_j \leq \underline{x}_j) = 0,$$
$$\theta_1^{i,M,j^*}(x_j \geq \overline{x}_j) = 1.$$

If additionally $\underline{x}_j < \overline{x}_j$, the range of x_j over which multiple equilibria exist in the MF market, $x_j \in (\underline{x}_j, \overline{x}_j)$ is non-empty.

*The **lower dominance region** in which a MF run is the unique equilibrium outcome irrespective of agents beliefs regarding other patient investors' actions is defined by:*

$$\underline{x}_j = \frac{c_j(\psi\bar{e}^M + (1-\psi)\kappa_0^{M,j})^2}{(1-\psi)\kappa_0^{M,j}}. \quad \forall\ \bar{e}^M > 0. \tag{29}$$

In the special case in which $\bar{e}^M = 0$, $\underline{x}_j = 0$, the lower dominance region is empty.

*The **upper dominance region** in which no patient MF investors choose to sell any MF shares early irrespective of their beliefs of other patient MF investors' actions is defined by:*

$$\overline{x}_j = \frac{c_j(\kappa_0^{M,j} - \psi\eta)^2}{(1-\psi)\kappa_0^{M,j}}. \tag{30}$$

The relative importance of payoff complementarities in patient MF investors' allocation decision diminishes with x_j. As the fundamental improves, MFs' index price impact decreases, while the early liquidation penalty increases in absolute terms. Fund outflows at $t = 1$ do not impose significant costs on remaining investors when expected long-term returns are high.

Corollary 6 *The size of the run (lower dominance) region $(0, \underline{x}_j]$ is increasing in c_j, the index illiquidity, as well as in e^M, the mass of impatient MF investors.*

This result directly follows from the partial derivatives of equation 29. MFs with more illiquid portfolios and investors with high short-term liquidity needs are more prone to runs.

Corollary 7 *The size of the no-run (upper dominance) region* $[\overline{x}_j, \infty)$ *is decreasing in* c_j *as well as in* $\kappa_0^{M,j}$*, the total mass of MF investors.*

Corollary 8 *In the special case in which* $c_j = 0$ *and MFs do not have price impact in index markets,* $\underline{x}_j = \overline{x}_j = 0$*. The region of* x_j *over which a MF run is the unique or one possible equilibrium outcome is empty. There are no negative externalities among MF investors and MFs never occur in equilibrium.*

For $x_j \in (\underline{x}_j, \overline{x}_j)$, there exist MF multiple equilibria at $t = 1$. The equilibrium outcome depends on agents' beliefs regarding other patient investors' redemption strategy, $E[\Delta\kappa_1^{M,j}|x_j]$. To overcome this multiplicity I use the sunspot equilibrium selection technique. In the region $(\underline{x}_j, \overline{x}_j)$, the equilibrium is determined by an i.i.d. sunspot for every realization of x_j.

Assumption 5 *For any* $x_j \in (\underline{x}_j, \overline{x}_j)$*, investors beliefs are such that they expect all other patient MF investors to run versus not run with equal probability. Formally:*

$$\pi \equiv Prob(\Delta\kappa_1^{M,j} = e^M + l^M | \underline{x}_j < x_j < \overline{x}_j) = 0.5.$$

Consequently, $Prob(\Delta\kappa_1^{M,j} = e^M | \underline{x}_j < x_j < \overline{x}_j) = 1 - \pi = 0.5.$

I use this specification to maintain tractability. The model predictions are robust to different values for π.

3.5 Equilibrium allocations in the initial period

At $t = 0$, each investor invests in the fund type that maximizes her expected lifetime wealth:

$$\theta_0^{i,E,j} = \begin{cases} 1 & \text{if } E_0[w_T^i|\lambda_i, \theta_0^{i,E,j} = 1] \geq E_0[w_T^i|\lambda_i, \theta_0^{i,M,j} = 1] \\ 0 & \text{if } E_0[w_T^i|\lambda_i, \theta_0^{i,E,j} = 1] < E_0[w_T^i|\lambda_i, \theta_0^{i,M,j} = 1] \end{cases} \tag{31}$$

and

$$\theta_0^{i,M,j} = \begin{cases} 1 & \text{if } E_0[w_T^i|\lambda_i, \theta_0^{i,M,j} = 1] > E_0[w_T^i|\lambda_i, \theta_0^{i,E,j} = 1] \\ 0 & \text{if } E_0[w_T^i|\lambda_i, \theta_0^{i,M,j} = 1] \leq E_0[w_T^i|\lambda_i, \theta_0^{i,E,j} = 1]. \end{cases} \tag{32}$$

The allocation policy defined by equations 31 and 32 assumes that an investor who is indifferent between the ETF or MF invests her entire endowment in the ETF. I abstract from mixed strategies and only focus on pure-strategy equilibria. This simplification does not affect the model predictions because the mass of investors with any given λ_i is infinitely small.

Assumption 6 *The investor who is indifferent between allocating her portfolio to ETFs or MFs at $t = 0$ invests her entire endowment in ETFs.*

The goal of this paper is to solve for $\theta^{i,E,j}$ and $\theta^{i,M,j}$ as a function of λ_i. If all investors allocated their entire endowment to the ETF in equilibrium, irrespective of their liquidity risks, $\theta^{i,E,j} = 1$ and $\theta^{i,M,j} = 0\ \forall \lambda_i$, ETFs would drive MFs extinct.

Investor i's expected payoff from investing in the ETF at $t = 0$ in equation 31 is given by:

$$E_0[w_T^i|\lambda_i, \theta_0^{i,E,j} = 1] = \mu_j - \lambda_i E_0[e^E](c_j + \phi_j), \tag{33}$$

where $E_0[e^E]$ is the expected mass of impatient ETF investors at $t = 1$. Equation 33 accounts for the result that patient ETF investors never choose to redeem early. ETF investors portfolio reallocation decision is independent of the fundamental. Impatient ETF investors receive $P_1^{E,j} = x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j)$ at $t = 1$. Patient ETF investors receive $P_2^{E,j} = x_j$. They account for the distribution of the fundamental and other ETF investors' liquidity risks in their initial allocation decision at $t = 0$.

Investor i's expected payoff from investing in the MF at $t = 0$ in equation 32 is given by

$$\begin{aligned} E_0[w_T^i|\lambda_i, \theta_0^{i,M,j} = 1] &= \lambda_i E_0[w_T^i|\lambda_i, \theta^{i,M,j} = 1 \cup i = e] \\ &\quad + (1-\lambda_i) E_0[w_T^i|\lambda_i, \theta^{i,M,j} = 1 \cup i = l] \\ &= \underbrace{\lambda_i \psi \mu_j}_{i \text{ is impatient}} + \underbrace{(1-\lambda_i) R^f \psi \left(\int_0^{\underline{x}_j} x_j dx + \frac{1}{2} \int_{\underline{x}_j}^{\overline{x}_j} x_j dx \right)}_{i \text{ is patient x run equilibrium}} \\ &\quad + \underbrace{(1-\lambda_i) \left(\frac{1}{2} \int_{\underline{x}_j}^{\overline{x}_j} (x_j - \Delta_2^{M,j}) dx + \int_{\overline{x}_j}^{\infty} (x_j - \Delta_2^{M,j}) dx \right)}_{i \text{ is patient x no-run equilibrium}}. \end{aligned} \tag{34}$$

The expected lifetime MF payoff does not only depend on investors' own types and the fundamental, but additionally depends on other MF investors' types and decisions and the sunspot. Not only the MF's terminal tracking difference, $\Delta_2^{M,j}$, but also the size of the run region $(\underline{x}_j, \overline{x}_j)$ depends on the size of the MF as of $t = 1$ as well as the distribution of liquidity risks among fund investors.

Within an index segment j, all cross-sectional variation in investors' initial allocations is due to heterogeneity in λ_i because everyone has identical preferences and there is no asymmetric information regarding x_j. The model is characterized by a cut-off equilibrium: Investors self-select into fund types based on their idiosyncratic liquidity needs.

Lemma 5 *If $(1-\psi)\mu_j$ is small relative to $c_j + \phi_j$, so there is price impact in index markets, $c_j > 0$, APs face balance sheet capacity constraints, $\phi_j > 0$, and MF prices are imperfectly*

flexible over the short-term, $\psi \to 1$, an investor with $\lambda_i = 1$ always invests in the MF at $t = 0$, whereas an investor with $\lambda_i = 0$ always invests in the ETF:

$$\theta_0^{i,M,j^*} = \begin{cases} 0 \text{ if } \lambda_i = 0 \\ 1 \text{ if } \lambda_i = 1 \end{cases}$$

$$\theta_0^{i,E,j^*} = \begin{cases} 1 \text{ if } \lambda_i = 0 \\ 0 \text{ if } \lambda_i = 1 \end{cases}$$

Th intuition behind lemma 5 is simple: The MF payoff at $t = 1$ is larger than the ETF payoff, $P_1^{M,j} > P_1^{E,j}$, when funds experience outflows and markets are illiquid. Under the same conditions, the ETF payoff at $t = 2$ is larger than the MF payoff, $P_2^{M,j} > P_2^{E,j}$. On one side, an investor with $\lambda_i = 1$, who faces a certain liquidity need at $t = 1$, always prefers the MF over the ETF. This investor values the short-term liquidity insurance provided by MFs' guaranteed redemption at the fund NAV and wants to avoid the short-term mispricing risk in ETFs. Since she always liquidates all fund holdings at $t = 1$, she does not face any MF share dilution risk. On the other side, a definitive long-term investor ($\lambda_i = 0$) always prefers the ETF over the MF. This type of investor wants to avoid the potential long-term share dilution in MFs and is not affected by the ETF's short-term mispricing risk. By definition, as a long-term investor she never needs to sell fund shares at a time when markets overall are illiquid and the cost of liquidity provision as reflected in the ETF mispricing is large.

Moving along the liquidity risk distribution starting at $\lambda_i = 1$, as λ_i decreases, the weight investors place on the fund payoff at $t = 1$ decreases while the weight placed on the payoff at $t = 2$ increases. Investors start to value MFs' short-term liquidity insurance less and increasingly care about reducing long-term share dilution risks. Eventually, when an investor's liquidity risk crosses a threshold level, λ', her expected return from investing in the ETF begins to exceed her expected return from the MF.

Proposition 5 *Investors facing high liquidity risks or short-term investment horizons self-select into mutual funds. Investors facing low liquidity risks or long-term investment horizons self-select into ETFs:*

$$\exists \lambda' \text{ for which } \{\theta_0^{i,E,j^*}, \theta_0^{i,M,j^*}\} = \{1, 0\} \ \forall i \text{ with } \lambda_i \leq \lambda'$$
$$\text{and } \{\theta_0^{i,E,j^*}, \theta_0^{i,M,j^*}\} = \{0, 1\} \ \forall i \text{ with } \lambda_i > \lambda'.$$

Importantly, the ETF and MF payoff are both increasing over time. The relative fund payoffs change over time, giving rise to the cut-off equilibrium described in proposition 5. The ETF payoff increases more strongly between $t = 1$ and $t = 2$ because of its larger sensibility to market liquidity conditions. After a period ($t = 1$) characterized by less liquidity and substantial

liquidity-driven fund liquidations, ETF investors in the model earn higher expected returns. Because the MF shields investors from changes in liquidity conditions and flow-induced price impact over the short-term, it provides them with more stable payoffs over time.

Fund and index prices. At $t = 0$, the fund sponsors receive capital investments from investors and convert them into MF and ETF shares respectively at an exchange rate of one. Initially, both funds and the index have the same price because fund shares are created outright at $t = 0$. ETF mispricing and MF tracking difference are zero:

$$P_0^{M,j} = P_0^{E,j} = P_0^j.$$

I normalize $P_0^j = 1$.

Fund market shares. Following proposition 5, the equilibrium market shares of the ETF and MF at $t = 0$ are given by:

$$MktShare_0^{E,j^*} = \lambda'$$
$$MktShare_0^{M,j^*} = 1 - \lambda'$$

3.6 Comparison to pre-ETF fund market equilibrium

This model presumes the coexistence of ETFs and MFs in a given index market segment. Consistent with this assumption, in the data many of the most popular benchmark indices are tracked by an ETF and MF. This has not always been the case. The first U.S. index MFs was launched already in 1976 by Vanguard (Vanguard 500 Index Fund) whereas ETFs were introduced in 1993 with the launch of SPY. Until then, MFs were the only option for households to cheaply access diversified portfolios and less liquid market segments. These developments beg the question of how the availability of ETFs, in addition to MFs, affects index investors' payoffs as well as the overall characteristics of funds' investor base and their vulnerability to early redemptions.

In principle, there are three variations of the index fund market: (i) MF monopoly, (ii) ETF monopoly, (iii) coexistence of ETF and MF. I refer to the model's baseline equilibrium, in which ETFs and MFs coexist, as the competitive equilibrium. Importantly, in this equilibrium there exists only one ETF and on MF. They compete solely with each other based on their common index portfolio. Table 1 summarizes investors' payoffs under the different fund market structures.

Table 1 provides three main insights: First, investors at the ends of the liquidity risk distribution with $\lambda_i = 0$ and $\lambda_i = 1$ are indifferent between the competitive equilibrium and the equilibrium featuring the monopoly of their preferred fund type. Following lemma 5, they are

not affected by other investors' choices as long as the fund type tailored to their liquidity needs exists. Yet, in the MF (ETF) monopoly in which their preferred fund does not exist investors with $\lambda_i = 0$ ($\lambda_i = 1$) are worse off. Second, ETF investors overall receive higher payoffs in the competitive equilibrium than in the ETF monopoly. They benefit from ETFs' competition with MFs because of the associated reduction in short-term mispricing risk. Third, MF investors generally receive lower payoffs in the MF monopoly versus the competitive equilibrium as they depend on the presence of other long-term investors for liquidity co-insurance.

Table 1: Investors' expected payoffs across different fund market structures

	$\lambda_i = 1$	$\lambda_i > \lambda'$	$\lambda_i \leq \lambda'$	$\lambda_i = 0$
CE	$\psi\mu_j$	$E[w^i_{T_i}\|\lambda_i, \theta^{i,M} = 1]$	$\mu_j - \lambda_i \Delta\kappa_1^{E,j}(c_j + \phi_j)$	μ_j
MF only	$\psi\mu_j$	$E[w^i_{T_i}\|\lambda_i, \theta^{i,M} = 1]$	$E[w^i_{T_i}\|\lambda_i, \theta^{i,M} = 1]$	$E[w^i_{T_i}\|\lambda_i, \theta^{i,M} = 1]$
ETF only	$\mu_j - \frac{1}{2}(c_j + \phi_j)$	$\mu_j - \lambda_i \frac{1}{2}(c_j + \phi_j)$	$\mu_j - \lambda_i \frac{1}{2}(c_j + \phi_j)$	μ_j

Note: The table shows the expected fund payoffs of investors, characterized by λ_i, under various fund market structures. CE refers to the competitive equilibrium in which investors choose between an ETF or MF. Expected CE payoffs are derived under investors' optimal investment policy following proposition 5. $E[w^i_{T_i}|\lambda_i, \theta^{i,M} = 1]$ is derived in equation 34. In the MF (ETF) monopoly, all payoffs are conditional on investments in the MF (ETF). ETF payoffs are in red and MF payoffs are in blue.

MF monopoly. The MF monopoly represents a pooling equilibrium in which all investors invest in the mutual fund irrespective of their λ_i. The law of large numbers implies that the mass of early MF investors in this setting is $\bar{e} = 0.5$. In expectation, half of the investors need liquidity at $t = 1$, while the other half are patient. Runs are possible. The run regions continue to be defined by equations 29 and 30.

On one hand, in the MF monopoly, the region in which a run is the unique equilibrium outcome is smaller than in the competitive equilibrium:

$$\underline{x}_j^{MF} > \underline{x}_j^{CE}. \tag{35}$$

This happens because the ratio of impatient to patient MF investors decreases relative to the competitive equilibrium. Intuitively, when relatively more long-term investors with $\lambda_i \leq \lambda'$, who would otherwise invest in ETFs, are pooled with more short-term investors ($\lambda_i > \lambda'$), the average horizon of the MF's investor base increases from $1.5 - \frac{\lambda'}{2}$ in the competitive equilibrium to 1.5. The average investment horizon in the MF monopoly is equal to the horizon of the average investor in the economy. As a result, the long-term share dilution cost caused by the impatient MF investors is shared among a larger group of patient MF investors.

On the other hand, in the MF monopoly, the region in which the no-run scenario is the unique equilibrium outcome is also smaller compared to the competitive equilibrium:

$$\overline{x}_j^{MF} < \overline{x}_j^{CE}. \tag{36}$$

As the fraction of impatient MF investors decreases, the share of patient MF investors increases. As a result, the potential share dilution costs associated with other patient MF investors' early liquidations rise. Anticipating this possibility, every individual patient MF investors faces stronger incentives to redeem her shares early if she believes other patient investors will do so as well.

Equations 35 and 36 together imply that the region over which multiple MF market equilibria are possible, $x_j \in (\underline{x}_j^{MF}, \overline{x}_j^{MF})$, is larger in the MF monopoly. Therefore, the effect of moving from the competitive equilibrium to the MF monopoly for overall run risk in this model is ambiguous as it depends on the exogenous sunspot probability of a run over the region defined by $x_j \in (\underline{x}_j^{MF}, \overline{x}_j^{MF})$.

At the investor level, long-term investors characterized by a low λ_i are worse off in the pooling MF monopoly compared to the competitive equilibrium. They have no choice but to subsidize the MF's short-term liquidity provision and are unlikely to benefit from the guaranteed MF redemption at NAV themselves. States in which they do redeem early are associated with a MF run. Investors on the other side of the liquidity risk spectrum, characterized by a high λ_i, are equally well off under both equilibria. They would choose to invest in MFs regardless and are likely to withdraw early. The medium λ_i investors benefit. These investors are able to access the MF's short-term liquidity insurance while sharing the long-term share dilution cost with the even more long-term investors.

ETF monopoly. The ETF monopoly represents a pooling equilibrium in which all investors invest in the ETF irrespective of their λ_i. Due to the absence of run risk in ETFs, in equilibrium half of the ETF investors liquidates their shares early, while the other, patient half waits until the terminal period. The total expected fund outflows in the interim period are strictly smaller in the ETF compared to the competitive equilibrium because inefficient early liquidation by patient investors never occur:

$$E[\Delta\kappa_1^{E,j,CE} + \Delta\kappa_1^{M,j,CE}] > E[\Delta\kappa_1^{E,j,ETF}] = 0.5.$$

Following 26, the ETF monopoly equilibrium prices are given by:

$$\begin{aligned} P_1^{E,j,ETF} &= x_j - \frac{1}{2}(c_j + \phi_j) < P_1^{E,j,CE}, \\ P_2^{E,j,ETF} &= x_j = P_2^{E,j,CE}. \end{aligned}$$

The long-term ETF payoff is unaffected by the larger mass of short-term ETF investors since every ETF investor bears her own transaction cost. The ETF price at $t = 1$ in the ETF monopoly is lower than in the competitive equilibrium. Intuitively, when investors' with $\lambda_i > \lambda'$ are invested in the ETF instead of the MF, there are more early ETF liquidations, $\Delta\kappa_1^{E,j,ETF} > \Delta\kappa_1^{E,j,CE}$. Amid market maker inventory costs and AP balance sheet capacity constraints, ETF investors' increased selling pressure leads to larger index market price impact and more ETF mispricing. These costs are born by the redeeming ETF investors. Investors facing the largest liquidity risks, $\lambda_i > \lambda'$, who would choose to invest in MFs if offered the option, are significantly worse off compared to the competitive equilibrium. Investors facing medium liquidity risks $0 < \lambda_i < \lambda'$ similarly suffer from the pooling of absence of short-term liquidity insurance in the ETF monopoly if they receive a liquidity shock at $t = 1$. Intuitively, all impatient ETF investors trade at the same equilibrium price. They do not only face the price impact associated with their trading activity, but also suffer from ETF mispricing. The larger the share of short-term ETF investors, the higher the ETF selling pressure, and therefore the more depressed the ETF market price. Only ETF investors with $\lambda_i = 0$ are truly indifferent between the ETF monopoly and the competitive equilibrium.

4 Policy implications

The model can be used to analyze several ongoing policy debates. I consider three applications – swing pricing, retirement plans and multi-fund share structures.

4.1 Swing pricing in open-end mutual funds

In November 2022, the SEC proposed a new rule to mandate the use of swing pricing for all U.S. open-end mutual funds other than money market funds and ETFs.[13] Swing pricing is a price-based liquidity management tool. It allows funds to adjust their NAV when faced with significant out- or inflows. For example, if the fund faces net outflows above a certain threshold (the swing threshold), it will reduce the NAV at which investors can redeem their shares by a pre-specified percentage factor (the swing factor). By imposing flow-induced trading costs on exiting or entering investors, swing pricing aims to mitigate payoff complementarities among MF investors. Consistent with the key MF friction in my model, the SEC highlighted the risk of flow-induced share dilution for MF shareholders as a main motivation for their proposed rule. Specifically, the SEC argues "Even when a fund manages its liquidity effectively, transaction costs associated with meeting redemption requests or investing the proceeds of subscriptions can create dilution for fund shareholders".[14]

I consider a model extension featuring full swing pricing. Under a full swing pricing policy, the MF price is "swung" whenever there are any net fund flows, $\Delta\kappa^{M,j} \neq 0$. In this case,

[13] File S7-26-22, Open-End Fund Liquidity Risk Management Programs and Swing Pricing; Form N-PORT.
[14] See https://www.sec.gov/files/33-11130-fact-sheet.pdf.

the swing threshold is effectively zero.[15] Accordingly, swing pricing simply imposes a wedge between the MF NAV and the price at which investors trade. Given a swing factor of $s_j > 0$, the MF price at at $t = 1$ becomes:

$$P_1^{M,j,Swing} = \begin{cases} NAV_1^{M,j} - s_j \text{ if } & \Delta\kappa_1^{M,j} > 0 \\ NAV_1^{M,j} + s_j \text{ if } & \Delta\kappa_1^{M,j} < 0. \end{cases} \tag{37}$$

In equilibrium the MF experiences net outflows in the interim period, $\Delta\kappa_1^{M,j} > 0$. Hence, the price received by MF investors at $t = 1$ under the swing pricing policy is lower than their payoff in the baseline model which is given by equation 23:

$$P_1^{M,j,Swing} < P_1^{M,j}.$$

Optimal swing factor. The main challenge in implementing swing pricing is the calibration of the swing factor s_j. In my model, the optimal swing factor is endogenously determined.

Proposition 6 *Within any index segment j and conditional on the volume of MF redemptions, $\Delta\kappa_1^{M,j}$, the optimal MF swing factor is given by:*

$$s_j{}^* = c_j \Delta\kappa_1^{M,j} \tag{38}$$

$s_j{}^*$ depends on the inventory cost of the market maker which is the root cause of flow-induced price impact in illiquid index markets in this framework. Adjusting the MF NAV by $s_j{}^*$ ensures that the MF's index market price impact at $t = 1$ is passed on to the leaving MF investors in proportion to their redemptions. If the MF price is set according to the pricing rule in equation 37 and the swing factor is calibrated in line with proposition 6, the price faced by MF investors at $t = 1$ is given by:

$$P_1^{M,j,Swing} = x_j - c_j \Delta\kappa_1^{M,j}.$$

In this case, $P_1^{M,j,Swing} = P_{1^+}^j$. Swing pricing makes the MF price perfectly forward looking. Then, corollaries 1 and 2 imply that the MF tracking difference, $\Delta_2^{M,J}$, must be zero under the optimal swing pricing rule. There is no MF share dilution risk, and therefore also no first-mover advantage.

Corollary 9 *Optimal swing pricing enables the MF to precisely replicate the index performance across all time horizons:*

[15] The alternative is partial swing pricing. Under partial swing pricing the swing threshold is non-zero.

$$P_1^{M,j,Swing} = P_{1^+}^j = x_j - c_j \Delta\kappa_1^{M,j},$$
$$P_2^{M,j,Swing} = P_2^j = x_j.$$

MF liquidity transformation is frictionless. The MF strictly dominates the same-index ETFs in terms of liquidity provision. In equilibrium, all rational investors choose to invest in the swing-pricing MF over the ETF:

$$\theta_0^{i,M,j,Swing} = 1 \quad \forall i$$
$$\theta_0^{i,E,j,Swing} = 0 \quad \forall i$$

Corollary 10 *If the MF utilizes swing pricing, the swing factor is calibrated optimally according to equation 38, and investors are rational and forward-looking, swing pricing eliminates any first-mover advantage within MFs. In equilibrium only impatient MF investors choose to redeem their shares early at $t = 1$:*

$$\Delta\kappa_1^{M,j,Swing} = \int_{i=e} \theta_0^{i,M,j} di$$

In terms of the cut-off equilibrium, corollary 9 implies:

$$\lambda'^{Swing} = 0.$$

With optimal swing pricing, the ETF and MF yield identical returns over the long term. In the short term, MFs outperform ETFs when the latter experience relative mispricing due to intermediary balance sheet constraints and both trade at the same index price. In this scenario, swing pricing does not only help MFs compete with ETFs, it can drive ETFs extinct.

Suboptimal swing factor. In practice, the optimal calibration of swing factors is challenging. While funds can observe the flow-based component of ${s_j}^*$, the price impact parameter, c_j, is more difficult to estimate accurately.[16] Unlike in this model, security market price impact may not always be linear in fund flows. Especially in periods of extreme market stress, trading costs can rise significantly. Price impact estimates derived from historical data may prove insufficient. Nonetheless, especially in these bad states (low x_j), the accurate calculation of swing factors is particularly important to minimize share dilution

[16] Currently, many MFs in the U.S. do not observe fund flows before they set their end-of-day NAV. This is an institutional constraint that could be eliminated by establishing a hard close on the time by which MF orders must be received in order to be executed at the day's NAV. Specifically, the MF, its transfer agent, or a registered clearing agency would need to receive the order information before the daily NAV is set. The latter typically happens at 4pm ET. The hard close requirement is part of the SEC's swing pricing proposal.

and consequently reduce run risks.Prior studies provide evidence in favor of the insufficient calibration of swing factors in jurisdictions where swing pricing is already commonly used (Jin, Kacperczyk, Kahraman, and Suntheim 2021; International Monetary Fund 2022).

If swing factors fail to comprehensively account for all the costs associated with investor flows, $s_j < c_j \Delta\kappa_1^{M,j}$, same-index ETFs and MFs continue to coexist in the model equilibrium. Swing pricing curbs share dilution for long-term MF investors by reducing short-term liquidity provision. Share dilution decreases but persists because payoffs to short-term MF investors continue to exceed their "fair share" of the fund's assets. Consequently, the marginal investor, characterized by her indifference between investing in the index MF and ETF in the no-swing pricing equilibrium, now prefers the MF over the ETF. The average expected investment horizon of the MF investor base increases. This reduces relative MF outflows as well as the risk of MF runs during states of market illiquidity.

Table 2 summarizes the fund payoffs under different calibrations of the MF swing pricing policy. Overall, within any specific index fund category, investors facing higher short-term liquidity risks are worse off following the implementation of swing pricing, regardless of its calibration. By design, swing pricing redistributes payoffs from investors who are trading to those who remain invested in the fund. If they had a choice, investors characterized by a high λ_i would always choose the no-swing pricing index MF. Conditional on swing pricing, they prefer suboptimally calibrated swing factors. Investors with "medium" liquidity needs (λ_i) distinctly benefit from a swing pricing policy. Under optimal swing pricing, the MF provides them with higher payoffs than the ETF in case they need liquidity at short notice while offering identical long-term returns. Only definitive long-term investors, $\lambda_i = 0$, remain indifferent between the ETF and MF. Under suboptimal swing pricing, the MF offers superior expected payoffs to investors located at the boundary between ETFs and MFs, $\lambda_i = \lambda'$, in the baseline model. Yet, investors who are less likely to need access to their capital at short notice continue to prefer ETFs over same-index MFs.

Table 2: Fund payoffs under different swing pricing policies

Swing factor	$P_1^{M,j}$	$P_1^{E,j}$	$P_2^{M,j}$	$P_2^{E,j}$
$s_j = 0$	$NAV_1^{M,j}$	$x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j)$	$x_j - \Delta_2^{M,j}$	x_j
$0 < s_j < s_j^*$	$x_j - s_j$	$x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j)$	$x_j - \tilde{\Delta}_2^{M,j}$	x_j
$s_j = s_j^*$	$x_j - \Delta\kappa_1^{M,j} c_j$	$x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j)$	x_j	x_j

Note: The table shows the fund payoffs at different times, $t = 1, 2$, and under various MF swing pricing policies. $s_j = 0$ refers to the baseline model specification without swing pricing, $s_j = s_j^*$ denotes swing pricing with optimally calibrated swing factors and $0 < s_j < s_j^*$ is swing pricing with swing factors that fall short of fully capturing the flow-induced trading costs in MFs.

Link to ETF prices. In this model, the optimal swing factor directly follows from the equilibrium ETF price in the interim period 26. The ETF price is determined intraday by market clearing between ETF investors and the AP. It reflects any price impact caused by AP arbitrage because the profit-maximizing AP simply passes on all costs associated with ETF creations or redemptions to the investors seeking liquidity in ETF markets. Yet, the ETF price is excessively flexible because it also reflects AP balance sheet capacity constraints. In the short-run, limits to ETF arbitrage imply that the law of one price does not necessarily hold with respect to ETF and index markets. The difference between the NAV and market price of an ETF tracking index j can reflect forward-looking price discovery and mispricing:

$$NAV_1^{E,j} - P_1^{E,j} = x_j - (x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j)) = \underbrace{\Delta\kappa_1^{E,j} c_j}_{\text{Price discovery}} + \underbrace{\Delta\kappa_1^{E,j}\phi_j}_{\epsilon_1^{E,j},\ \text{mispricing}} \tag{39}$$

Corollary 11 *When APs face balance sheet capacity constraints, $\phi_j > 0$, ETF discounts represent an upper bound on the optimal swing factor:*

$$\frac{NAV_1^{E,j} - P_1^{E,j}}{\Delta\kappa_1^{E,j}} > \frac{{s_j}^*}{\Delta\kappa_1^{M,j}}.$$

Anadu, Levin, Liu, Tanner, Malfroy-Camine, and Baker (2023) suggest to use observed ETF discounts as a proxy to calibrate swing factors for MFs holding similar portfolios. Equation 39 and corollary 11 illustrate that in states where APs are balance sheet constrained and ETF arbitrage is imperfect, swing factors derived from ETF discounts may lead to an overestimation of the trading costs associated with the redemption of MF investors. This, in turn, can result in excessive liquidation costs being imposed on MF investors seeking redemption.

4.2 Index investments within retirement accounts

Although, I model portfolio allocations made by investors outside of their retirement investment accounts, this study applies to index fund selection within retirement accounts, an area of paramount significance. Until they near their retirement, retirement investors are long-term investors who face significant penalties for early withdrawals. Retirement assets invested in index MFs are usually commingled with those originating from non-retirement accounts offering daily liquidity. It is noteworthy that distinct share classes, catering to retirement, institutional, and retail investors, frequently coexist within a single MF, each sharing the same underlying portfolios and associated trading costs. Investors holding index MFs with illiquid portfolios within their retirement accounts may inadvertently subsidize the short-term liquidity provision to non-retirement account investors. This cross-subsidization effect becomes particularly relevant when a substantial portion of the fund's equity is held by investors who exhibit a recurring demand for liquidity or a propensity to divest their holdings when market liquidity is low. In contrast, I demonstrate that the long-term payoff of index

ETFs remains unaffected by other investors' intraday trading activities.

An additional cost of investing in ETFs as compared to no-load open-end MFs are bid-ask spreads. Similar to the possibility of relative mispricing in ETFs, bid-ask spreads represent a cost associated with providing short-term liquidity. While bid-ask spreads for broadly diversified index ETFs typically remain modest during stable market conditions, it is noteworthy that these spreads may accrue over time, potentially affecting the overall cost of ETF investments and making them less attractive within retirement accounts. The economic magnitude of these effects is an empirical question and should be considered in future research on the potential role of ETFs in retirement accounts.

4.3 ETF share classes in open-end Mutual Funds

Another model application is the analysis of investor trade offs within dual fund share structure. In this structure the ETF is a share class of an existing MF portfolio instead of a standalone fund. Vanguard is currently the only U.S. asset manager using this structure, managing around $2 trillion of index ETF assets, equivalent to nearly 30% of all U.S. ETF assets.[17] The rationale behind this fund structure is that if investors expected to trade frequently, they could invest in, or convert their MF shares into the ETF share class and sell on the exchange without impacting any other fund investors. As a case in point Vanguard's former Chief Investment Officer George U. "Gus" Sauter explained the dual fund structure by saying "I started thinking, if we had a share class of the funds that traded on the exchange, then likely the people with a shorter time horizon would migrate to that share class. And if they decided they needed to get out of the market, they could sell off their stock in that share class, as opposed to trying to come out of the traditional mutual fund share class."[18] Since the expiry of its patent protection in May 2023, several other asset managers have submitted applications to the SEC to copy this structure.[19] Thus, it is important to understand the implications of joint ETF-MF fund structures for investor payoffs and redemption incentives.

Abstracting from fee differences, in the joint ETF-MF fund structure the MF and ETF share classes have the same NAV. ETFs are traded at the market price which may differ from the fund NAV. As with multiple MF share classes, any trading costs are shared across all share classes, including the ETF share class, on a pro rata basis. As usual, only MF shares are traded at NAV. Holding constant the investor composition among funds, over the short-term at $t = 1$ the ETF and MF payoffs are the same as in the baseline model:

[17] Morningstar data as of May 31, 2023. The structure has been patented under the U.S. patent no.: US 6,879,964 B2 and been granted exempted relief from the 1940 Investment Company Act by the SEC.

[18] See https://www.ft.com/content/e57771ce-bdcd-4ab2-ba8b-cb472279f366.

[19] See for example the application by Dimensional Fund Advisors available at https://www.sec.gov/Archives/edgar/data/354204/000168035923000216/dimensional40app07122023.htm.

$$P_1^{M,j,Dual} = NAV_1^{M,j},$$
$$P_1^{E,j,Dual} = x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j) = P_1^{E,j}.$$

$\kappa_1^{E,j,Dual}$ refers to the sell volume of the funds' ETF shares on exchange. The fund NAV is determined through the same mechanisms as for a standalone MF structure, so it equals the value of the fund's index holdings divided by the sum of its MF and ETF shares outstanding:

$$NAV_2^{M,j,Dual} = NAV_2^{E,j,Dual} = \underbrace{x_j}_{P_2^j} \underbrace{(1 - \frac{\Delta\kappa_1^{M,j} P_1^{M,j}}{P_{t^+}^j} - \kappa_1^{E,j})}_{\substack{\text{\# index shares remaining} \\ \text{in joint fund portfolio}}} * \underbrace{(\kappa_1^{M,j} + \kappa_1^{E,j})^{-1}}_{\substack{\text{MF + ETF} \\ \text{shares} \\ \text{outstanding}}}. \quad (40)$$

It follows for the terminal ETF and MF payoff under the dual fund structure:

$$P_2^{M,j,Dual} = NAV_2^{M,j,Dual} > P_2^{M,j}$$
$$P_2^{E,j,Dual} = NAV_2^{E,j,Dual} < \underbrace{P_2^{E,j}}_{x_j}.$$

MF investors get a better deal under the dual fund structure. ETF investors are worse off. The relative payoffs for early MF investors stay unchanged but the long-term investors' payoffs increase. While the MF share class remains unaffected by ETF-specific mispricing risk, the ETF share class inherits share dilution risks from the MF. For MF investors the arrangement implies that share dilution costs caused by redemptions from MF share classes are shared with investors in the ETF share class. This is evident from a comparison of equation 40 and the terminal MF NAV of the standalone MF in equation 15. As a result, the expected liquidity insurance cost for MF investors decreases. As long as some ETF investors remain invested until $t = 2$, $\kappa_1^{E,j} > 0$, the MF tracking difference per share is lower under the dual fund structure compared to the baseline model, $\Delta_2^{M,j,Dual} < \Delta_2^{M,j,Dual}$. Higher long-term payoffs for MF shareholders in a multi-fund structure represent a redistribution from ETF investors. Patient ETF investors subsidize the MF's short-term liquidity provision. The specific magnitude of this liquidity cross-substitution depends on the amount of early redemptions from the MF share classes and the illiquidity of fund assets at the time of the redemptions, c_j. Early ETF investors' payoff is the same as in a standalone ETF since it only depends on the index value, market liquidity, and intermediary balance sheet capacity. These factors do not change when an ETF shares its portfolio with MF share classes.

In equilibrium, rational investors with $\lambda_i > 0$ never choose to invest in the ETF structured as a share class of an existing MF. The ETF is dominated by the MF share class:

$$P_1^{E,j,Dual} < P_1^{M,j,Dual},$$
$$P_2^{E,j,Dual} = P_2^{M,j,Dual}.$$

These findings corroborate the SEC's concerns about cross-subsidization in multi-fund share structures, as articulated in its 2019 ETF rule. They also bolster its reluctance to authorize additional fund companies to merge ETF and MF share classes within a single fund.[20]

5 Empirical predictions

I focus on the two-fund equilibrium characterized by the coexistence of ETFs and MFs to derive predictions about the cross-section of investor portfolio allocations. The key model predictions relate to investors' portfolio choice between same-index ETFs and MFs as a function of their individual, heterogeneous liquidity needs. Moreover, my model generates predictions on the impact of changes in the cross-sectional distribution of liquidity needs on the relative size of ETFs and MFs. To conclude, I explore predictions for a fund market equilibrium scenario wherein MFs implement swing pricing.

Prediction 1 *In less liquid market segments, long-term investors or investors facing minimal liquidity risks invest in ETFs.*

Prediction 2 *In less liquid market segments, investors with shorter investment horizons or greater liquidity needs allocate a larger portion of their portfolio to open-end mutual funds.*

Predictions 1 and 2 directly follow from proposition 5. In the competitive equilibrium investors characterized by $\lambda_i > \lambda'$ invest in the MF, while those with $\lambda_i \leq \lambda'$ opt for the ETF. Investors may invest different parts of their fund portfolio for distinct purposes. For example, they may earmark some assets to cover unexpected expenses, financial emergencies, or to provide short-term liquidity in case of job loss. Concurrently, they invest to accumulate savings for long-term objectives, such as retirement. Consequently, an alternative interpretation of predictions 1 - 2 posits that rational investors place the portion of their portfolio designated for short-term liquidity needs in MFs, reserving ETFs for long-term savings.

The ideal setting for testing predictions 1 and 2 involves a panel of detailed, investor-level fund holdings data, including information on investors' cash flows (e.g., labor income and expenses), and overall financial wealth. Investor demographic, income and wealth information is required for estimating each in individual's distinct liquidity needs. Crucially, not only investor-level data but also fund-level information is required. This is essential for distinguishing between ETF and MF holdings as well as for controlling fund-specific features, such as the benchmark index and fees. Relying solely on aggregate data for an investors' total

[20] In the final rule the SEC states, "For example, an ETF share class that transacts with authorized participants on an in-kind basis and a mutual fund share class that transacts with shareholders on a cash basis may give rise to differing costs to the portfolio. As a result, while certain of these costs may result from the features of one share class or another, all shareholders would generally bear these portfolio costs." See Exchange-Traded Funds Release No. IC-33646 (Sept 25, 2019) p. 122-123.

ETF and open-end mutual fund holdings proves inadequate, as the model generates distinct predictions for investors' fund allocations in liquid versus less liquid market segments.

Prediction 3 *In index segments (asset classes) where the overall investor base exhibits relatively larger short-term liquidity needs (shorter holding periods), mutual funds dominate with a larger market share compared to ETFs.*

Prediction 4 *In index segments (asset classes) where the majority of investors pursue long-term investment strategies, ETFs hold a larger market share compared to mutual funds.*

Predictions 3 and 4 pertain to the long-term equilibrium fund market composition and stem from proposition 5 together with the cross-sectional distribution of investors' liquidity needs, denoted as $\lambda_i \sim U[\lambda_0, \lambda_1]$. A higher value for $\lambda_0 \geq 0$ signifies greater average liquidity needs across the investor spectrum, while a lower value for $\lambda_1 \leq 1$ indicates that the average investor is less likely to be impatient. Changes in these distributional parameters impact the expected proportion of impatient investors, thereby affecting the equilibrium sizes of MFs and ETFs.

Prediction 5 *In less liquid market segments, MFs implementing swing pricing, with swing factors calibrated to approximate but not surpass the trading costs induced by fund flows, attract longer-term investors and are characterized by larger average holding periods than otherwise identical MFs without swing pricing.*

Prediction 6 *In less liquid market segments, the adoption of swing pricing increases MFs' market share relative to ETFs.*

Predictions 5 and 6 are based on proposition 6. When MFs adopt swing pricing, they become more attractive to long-term investors who continue to face some residual short-term liquidity risks. As long as the swing factors applied do not exceed transaction costs associated with funds flow-induced trading activity, short-term investors continue to prefer MFs over ETFs due to the latter's susceptibility to mispricing risks. As of the current date, the SEC has not finalized its proposed swing pricing rule, and no U.S. MFs, excluding money market mutual funds, currently employ swing pricing. Consequently, predictions 5 and 6 are not yet testable empirically in U.S. data. Yet, should the SEC proceed with its proposal, or if regulators in other jurisdictions enhance data availability concerning the utilization of swing pricing across funds, these predictions can be examined empirically in the future.

6 Conclusion

I study rational investors' optimal portfolio allocations between ETFs and index open-end mutual funds. This paper's main message is that the distinct trading and pricing mechanisms underlying ETFs and mutual funds matter for relative liquidity provision at different investment horizons. Although same-index ETFs and MFs share identical fundamental risks

through their common security portfolios, their payoffs may not always be equivalent. When investors face heterogeneous liquidity needs, ETFs are not universally preferred over same-index mutual funds by all investors. Instead, both fund types can co-exist in an index fund market equilibrium because they cater to investors' varied liquidity needs.

This result follows directly from the distinct mechanisms through which the costs of the liquidity transformation provided by these investment funds are reflected in their share prices. ETFs and MFs represent alternative technologies designed to provide investors' with access to inherently illiquid market segments and cheap diversification. MFs are an intermediary-based index investing technology as part of which investors co-insure each other against short-term liquidity needs through the guaranteed redemption of fund shares at the end-of-day fund NAV. Particularly in illiquid index segments, such as corporate bond or international equity funds, and during periods of market stress, MFs' unique payoff structure shields investors with urgent liquidity needs from potentially high trading costs. The costs of MFs' short-term liquidity provision manifest as share dilution and are borne by remaining, long-term MF investors. The resulting payoff complementarities among MF investors give rise to run risks, the key friction associated with this technology.

Conversely, ETFs constitute a market-based investing technology. Liquidity provision occurs via intermediaries at two levels, in security and ETF markets. There is no co-insurance between fund investors since each ETF investor bears her own transaction costs when trading ETF shares in secondary markets. This forces ETF investors to internalize their price impact in illiquid security markets and limits the potential for flow-induced share dilution. Yet, because the pricing of ETF shares depends on the continuous secondary-market liquidity provision by APs, over the short term, AP balance sheet constraints may cause disparities between ETF prices and their underlying portfolio NAV, resulting in excessive price volatility. ETF mispricing represents the liquidity premium earned by APs in exchange for their ETF creation and redemption activities. Over the long term, the law of one price generally prevails, and once AP balance sheet constraints subside, ETF prices tend to converge to the fund NAV.

In this framework, MFs are naturally preferred by investors with relatively higher liquidity needs or shorter investment horizons. These investors are willing to sacrifice long-term expected returns in exchange for MFs' short-term liquidity insurance. ETFs in turn are preferred by investors with lower liquidity needs and longer-term investment horizons. ETFs may be more suited for less liquid index market segments favored by long-term investors, whereas MFs may be a better fit in liquid fund market segments favored by investors with short-term liquidity needs, such as money market funds. Both funds are virtually perfect substitutes in highly liquid market segments, such as large-cap domestic equities.

These findings hold significant implications for policy makers. First, the SEC's proposed

swing pricing rule could benefit MFs by reducing share dilution costs tied to early redemptions, attracting more long-term investors, and thereby allowing funds to maintain a larger size. Second, regulators should encourage retirement plan sponsors to include ETFs as part of their investment options. This would enable retirement investors to adjust their portfolio allocation between ETFs and MFs, aligning their investments more effectively with their liquidity needs and investment horizons, both before and during retirement. Third, the SEC should not permit multi-fund structures featuring ETF and MF share classes within one fund in less liquid market segments, as such set-ups benefit MF shareholders at the cost of ETF investors.

My study opens many avenues for future empirical and theoretical research. It raises questions such as: Does the introduction of ETFs decrease staleness in MF NAVs? How do conflicts of interest among financial intermediaries acting as APs for ETFs and market makers for MFs affect trade execution costs for mutual funds? Additionally, what financial stability risks arise from the trend towards third-party ETF arbitrage, where APs conduct ETF creations or redemptions on behalf of proprietary trading firms?

References

Agapova, A. (2011). Conventional Mutual Index Funds Versus Exchange Traded Funds. *Journal of Financial Markets 14*(2), 323–343.

Agarwal, V., H. Ren, K. Shen, and H. Zhao (2022). Redemption in kind and mutual fund liquidity management. Working paper.

Anadu, K., J. Levin, V. Liu, N. Tanner, A. Malfroy-Camine, and S. Baker (2023). Swing Pricing Calibration: Using ETFs to Infer Swing Factors for Mutual Funds. *Financial Analysts Journal 0*(0), 1–11.

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

Ben-David, I., F. Franzoni, and R. Moussawi (2018). Do ETFs Increase Volatility? *The Journal of Finance 73*(6), 2471–2535.

Bhattacharya, A. and M. O'Hara (2018). Can ETFs Increase Market Fragility? Effect of Information Linkages in ETF Markets. Working paper.

Boudoukh, J., M. Richardson, M. Subrahmanyam, and R. F. Withelaw (2002). Stale Prices and Strategies for Trading Mutual Funds. *Financial Analyst Journal 58*(4), 53–71.

Box, T., R. L. Davis, and K. P. Fuller (2019). ETF Competition and Market Quality. *Financial Management 48*(3), 873–916.

Brown, C., S. Davies, and M. Ringgenberg (2021). ETF arbitrage, non-fundamental demand, and return predictability. *Review of Finance 25*(4), 937–972.

Cass, D. and K. Shell (1983). Do Sunspots Matter? *Journal of Political Economy 91*(2), 193–227.

Chalmers, J. M. R., R. M. Edelen, and G. B. Kadlec (2001). On the Perils of Financial Intermediaries Setting Security Prices: The Mutual Fund Wild Card Option. *The Journal of Finance 56*(6), 2209–2236.

Chen, Q., I. Goldstein, and W. Jiang (2010, August). Payoff complementarities and financial fragility: Evidence from mutual fund outflows. *Journal of Financial Economics 97*(2), 239–262.

Cherkes, M., J. Sagi, and R. Stanton (2009). A liquidity-based theory of closed-end funds. *Review of Financial Studies 22*(1), 257–297.

Choi, J., M. Kronlund, and J. Y. J. Oh (2022). Sitting bucks: Stale pricing in fixed income funds. *Journal of Financial Economics 145*(2, Part A), 296–317.

Choi, J., D. Laibson, and B. C. Madrian (2009, 11). Why Does the Law of One Price Fail? An Experiment on Index Mutual Funds. *The Review of Financial Studies 23*(4), 1405–1432.

Chordia, T. (1996). The structure of mutual fund charges. *Journal of Financial Economics 41*(1), 3–39.

Coval, J. and E. Stafford (2007). Asset fire sales (and purchases) in equity markets. *The Journal of Financial Economics 86*, 479–512.

Da, Z. and S. Shive (2018). Exchange traded funds and asset return correlations. *European Financial Management 24*(1), 136–168.

Dannhauser, C. and S. Hoseinzade (2017). The transformation of corporate bond investors and fragility: Evidence on mutual funds and etfs. Working paper.

Dávila, E. and I. Goldstein (2023). Optimal Deposit Insurance. *Journal of Political Economy 131*(7), 1676–1730.

Deli, D. N. and R. Varma (2002). Closed-end versus open-end: the choice of organizational form. *Journal of Corporate Finance 8*(1), 1–27.

Diamond, D. W. and P. H. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy 91*(3), 401–419.

Dickson, J. M., J. B. Shoven, and C. Sialm (1999). Tax Externalities of Equity Mutual Funds. *National Tax Journal 53*(93).

Edelen, R. M. (1999). Investor flows and the assessed performance of open-end mutual funds. *Journal of Financial Economics 53*(3), 439–466.

Elton, E. J., M. J. Gruber, C. R. Blake, and O. Shachar (2013). Why Do Closed-End Bond Funds Exist? An Additional Explanation for the Growth in Domestic Closed-End Bond Funds. *The Journal of Financial and Quantitative Analysis 48*(2), 405–425.

Elton, E. J., M. J. Gruber, and J. A. Busse (2004). Are Investors Rational? Choices among Index Funds. *The Journal of Finance 59*(1), 261–288.

Falato, A., I. Goldstein, and A. Hortaçsu (2021). Financial Fragility in the COVID-19 Crisis: The Case of Investment Funds in Corporate Bond Markets. *Review of Financial Economics 123*, 35–52.

Federal Reserve (2000). Mutual funds and the u.s. equity market. Federal reserve bulletin 2000, Federal Reserve.

Feroli, M., A. Kashyap, K. Schoenholtz, and H. S. Shin (2014). Market Tantrums and Monetary Policy. Report for the 2014 u.s. monetary policy forum.

Giuzio, M., M. Grill, D. Kryczka, and C. Weistroffer (2021). A Theoretical Model Analysing Investment Funds' Liquidity Management and Policy Measures. Macroprudential Bulletin 12, European Central Bank.

Glosten, L., S. Nallareddy, and Y. Zou (2021). ETF Activity and Informational Efficiency of Underlying Securities. *Management Science 67*(1), 22–47.

Goetzmann, W. N., Z. Ivković, and K. G. Rouwenhorst (2001). Day Trading International Mutual Funds: Evidence and Policy Solutions. *The Journal of Financial and Quantitative Analysis 36*(3), 287–309.

Goldstein, I., H. Jiang, and D. T. Ng (2017). Investor flows and fragility in corporate bond funds. *Journal of Financial Economics 126*(3), 592–613.

Goldstein, I. and A. Pauzner (2005). Demand–Deposit Contracts and the Probability of Bank Runs. *The Journal of Finance 60*(3), 1293–1327.

Gorbatikov, E. and T. Sikorskaya (2022). Two APs Are Better Than One: ETF Mispricing and Primary Market Participation. Working papaer.

Greene, J. T. and C. W. Hodges (2002). The dilution impact of daily fund flows on open-end mutual funds. *Journal of Financial Economics 65*(1), 131–158.

Gromb, D. and D. Vayanos (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics 66*(2-3), 361–407.

Haddad, V., A. Moreira, and T. Muir (2021, 01). When Selling Becomes Viral: Disruptions in Debt Markets in the COVID-19 Crisis and the Fed's Response. *The Review of Financial Studies 34*(11), 5309–5351.

Hortacsu, A. and C. Syverson (2004). Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of SandP 500 Index Funds. *The Quarterly Journal of Economics 119*(2), 403–456.

Huang, J. and I. Guedj (2009). Are ETFs replacing index mutual funds? Afa 2009 san francisco meetings paper.

ICI (2023). 2023 Investment Company Fact Book. Technical report.

International Monetary Fund (2022). Chapter 3 – Asset Price Fragility in Times of Stress: The Role of Open-End Investment Funds. Global financial stability report.

Israeli, D., C. Lee, and S. A. Sridharan (2017). Is there a dark side to exchange traded funds? An information perspective. *Review of Accounting Studies 22*(3), 1048–1083.

Jacklin, C. (1983). Demand Deposits, Trading Restrictions and Risk Sharing. Working paper.

Jiang, Hao, D. L. and A. Wang (2021). Dynamic Liquidity Management by Corporate Bond Mutual Funds. *Journal of Financial and Quantitative Analysis 56*(5), 1622–52.

Jin, D., M. Kacperczyk, B. Kahraman, and F. Suntheim (2021, 03). Swing Pricing and Fragility in Open-End Mutual Funds. *The Review of Financial Studies 35*(1), 1–50.

Johnson, W. T. (2004). Predictable Investment Horizons and Wealth Transfers among Mutual Fund Shareholders. *The Journal of Finance 59*(5), 1979–2012.

Kacperczyk, M. and P. Schnabl (2013). How Safe Are Money Market Funds? *The Quarterly Journal of Economics 128*(3), 1073–1122.

Khomyn, M., T. J. Putniņš, and M. Zoican (2023). The value of etf liquidity. Working paper.

Koont, N., Y. Ma, L. Pastor, and Y. Zeng (2023). Steering a Ship in Illiquid Waters: Active Management of Passive Funds. Working paper.

Lettau, M. and A. Madhavan (2018, February). Exchange-Traded Funds 101 for Economists. *Journal of Economic Perspectives 32*(1), 135–54.

Ma, Y., K. Xiao, and Y. Zeng (2022a). Bank Debt, Mutual Fund Equity, and Swing Pricing in Liquidity Provision. Working paper.

Ma, Y., K. Xiao, and Y. Zeng (2022b, October). Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity. *Review of Financial Studies 35*(10), 4674–711.

Madhavan, A. and D. Morillo (2018). The Impact of Flows into Exchange-Traded Funds: Volumes and Correlations. *The Journal of Portfolio Management 44*, 96–107.

Madhavan, A. and A. Sobczyk (2016). Price dynamics and liquidity of exchange-traded funds. *Journal of investment management 14*, 86–102.

Malamud, S. (2016). A Dynamic Equilibrium Model of ETFs. Discussion paper DP11469, CEPR.

Morris, S. and H. S. Shin (2003). *Global Games: Theory and Applications*. Cambridge: Cambridge University Press.

Moussawi, R., K. Shen, and R. Velthius (2022). The Role of Taxes in the Rise of ETFs. Working paper.

Pan, K. and Y. Zeng (2019). ETF Arbitrage Under Liquidity Mismatch. Working Paper 485311, Harvard University OpenScholar.

Petajisto, A. (2017). Inefficiencies in the Pricing of Exchange-Traded Funds. *Financial Analysts Journal 73*, 24–54.

Schmidt, L., A. Timmermann, and R. Wermers (2016). Runs on Money Market Mutual Funds. *American Economic Review 106*(9), 2625–2657.

Shim, J. J. (2019). Arbitrage Comovement. Working paper.

Shim, J. J. and K. Todorov (2022). ETFs, Illiquid Assets, and Fire Sales. Working paper.

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

Tufano, P., M. Quinn, and R. Taliaferro (2012). Live Prices and Stale Quantities: $T + 1$ Accounting and Mutual Fund Mispricing. *Journal of Investment Management 10*(1), 5–15.

Wermers, R. (2000). Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses. *Journal of Finance 55*(4), 1655–1695.

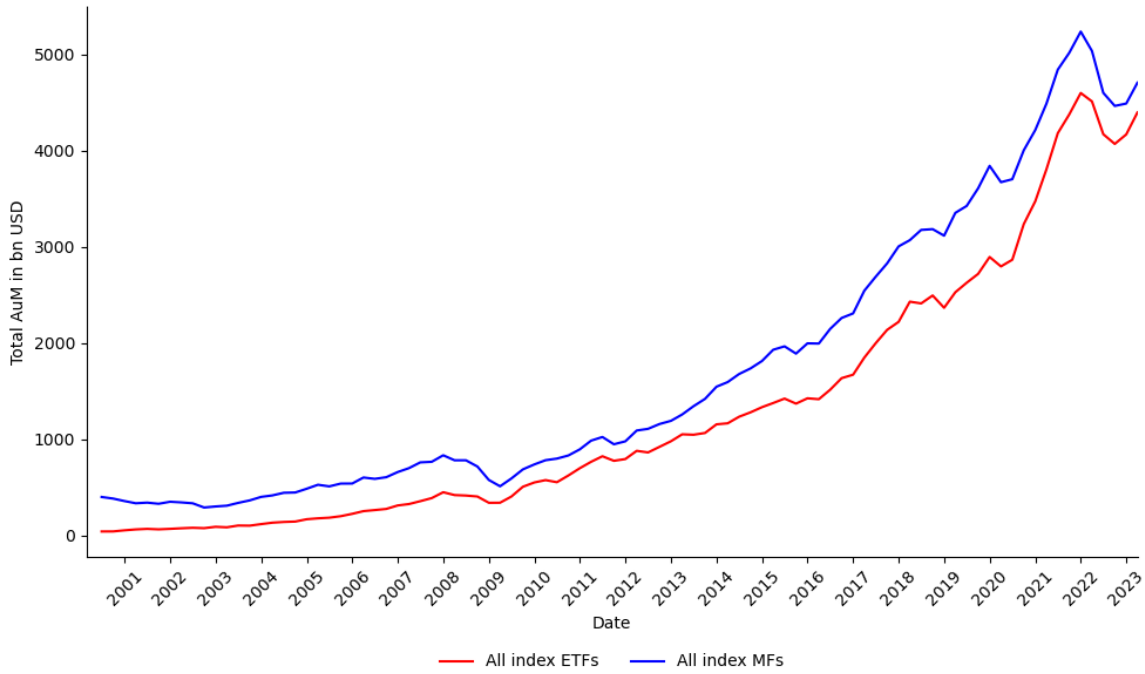
Zeng, Y. (2017). A Dynamic Theory of Mutual Fund Runs and Liquidity Management. Working paper.

Zitzewitz, E. (2003). Who Cares about Shareholders? Arbitrage-Proofing Mutual Funds. *Journal of Law, Economics, and Organization 19*(2), 245 – 280.

Appendices

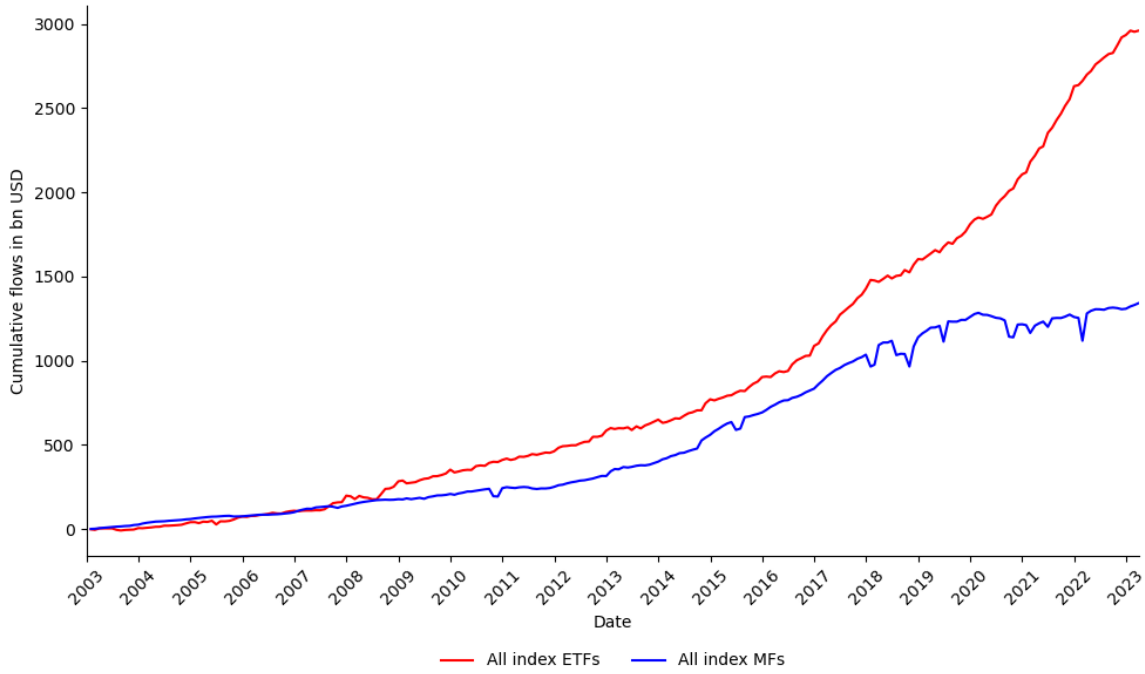
A Figures

Figure A.1: Size of U.S. Index Fund Market by Fund Type



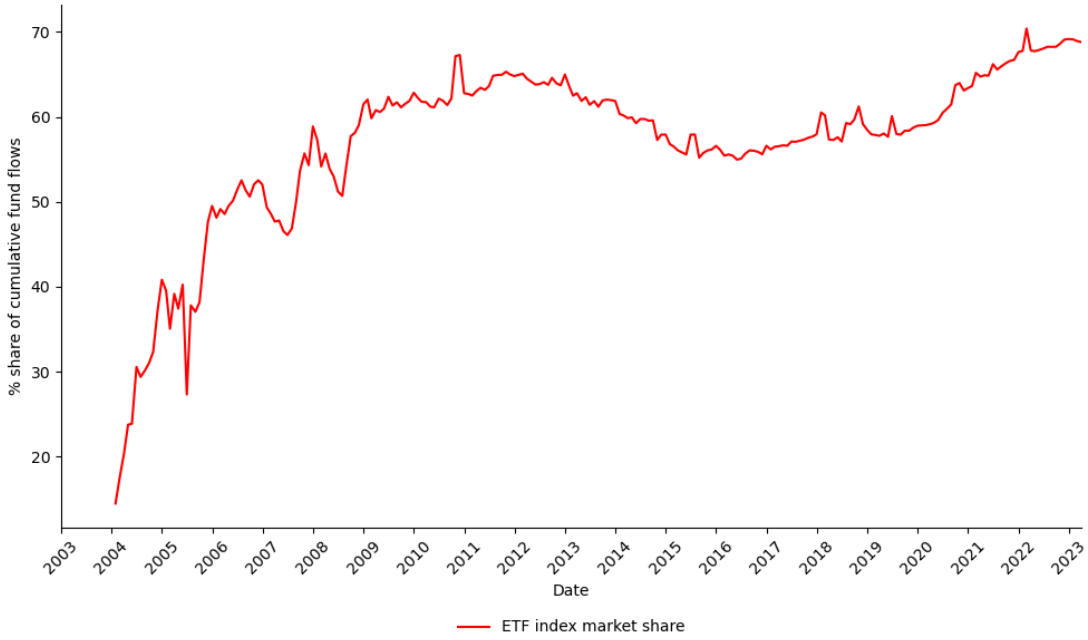
Note: This figure shows the AUM in the CRSP sample of U.S. based index ETFs and open-end mutual funds in bn USD. Sample funds include unit investment trusts but exclude closed-end funds, levered and inverse funds.

Figure A.2: Cumulative Flows into U.S. Index ETFs and MFs



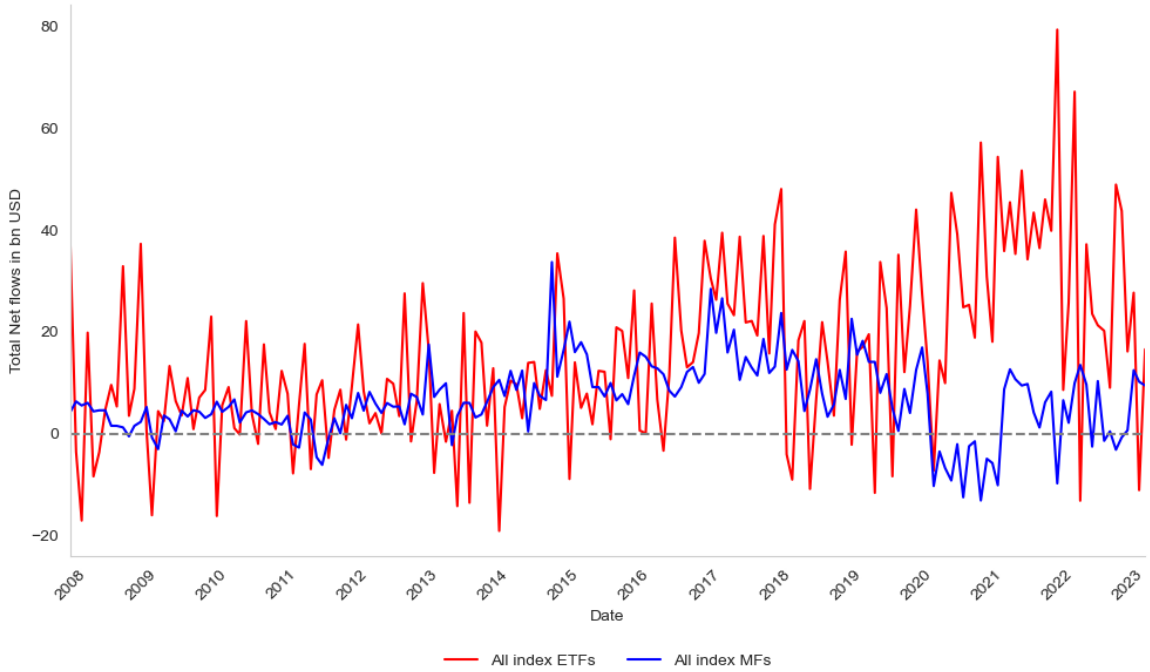
Note: This figure shows the cumulative monthly fund flows into U.S. based index ETFs and open-end mutual funds since January 2003 in bn USD. Cumulative flows are calculated as $Cum\ Flow_t^{ETF} = \sum_{t=1}^{T}(Shares\ Out_t - Shares\ Out_{t-1}) * NAV_t$ for ETFs and $Cum\ Flow_t^{MF} = \sum_{t=1}^{T} AUM_t - AUM_{t-1} * (1 + r_t^{NAV})$ for MFs, where r_t^{NAV} is the monthly NAV-based MF return. Estimated fund flows are winsorized at the 1^{st} and 99^{th} percentile before aggregation.

Figure A.3: ETF % Market Share in Cumulative U.S. Index Fund Flows



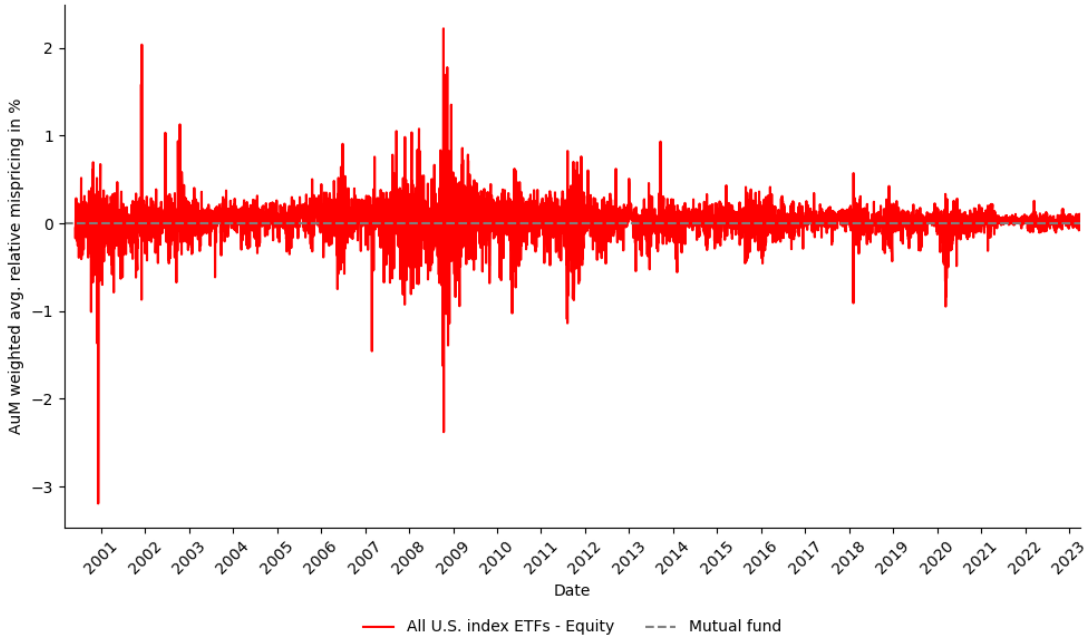
Note: This figure shows the flows into index ETFs in percent of the cumulative flows into U.S. based ETFs and open-end mutual funds.

Figure A.4: Monthly Flows into U.S. Index ETFs and MFs



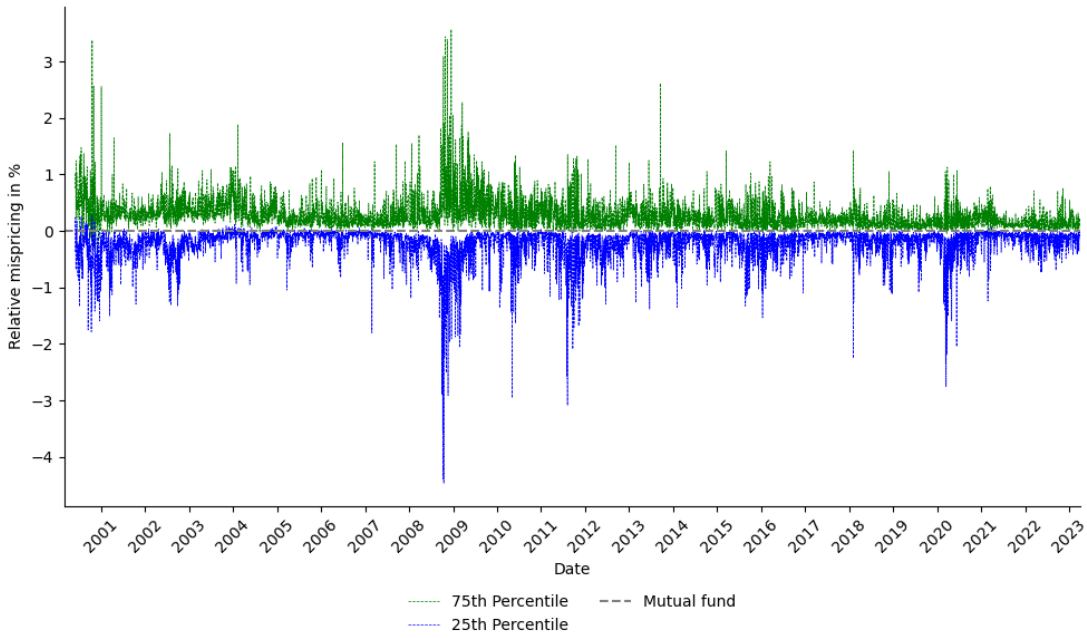
Note: This figure shows the monthly fund flows into U.S. based index ETFs and open-end mutual funds in bn USD. Monthly flows are calculated as $Flow_t^{ETF} = (Shares\ Out_t - Shares\ Out_{t-1}) * NAV_t$ for ETFs and $Flow_t^{MF} = AUM_t - AUM_{t-1} * (1 + r_t^{NAV})$ for MFs, where r_t^{NAV} is the monthly NAV-based MF return. Estimated fund flows are winsorized at the 1^{st} and 99^{th} percentile before aggregation.

Figure A.5: Asset-Weighted Relative Mispricing for Equity Index ETFs



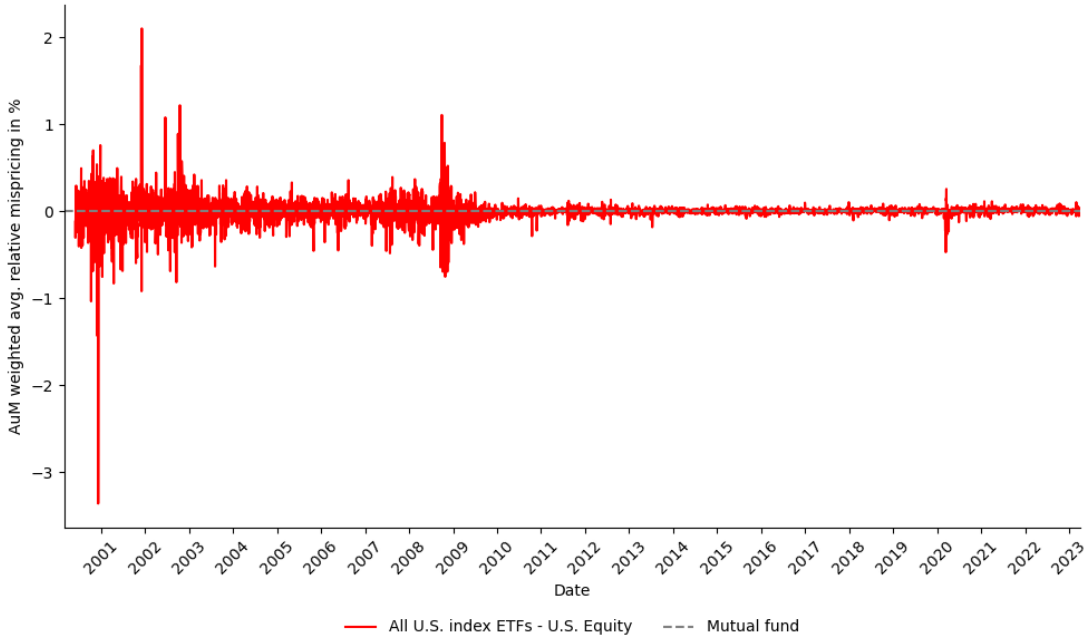
Note: This figure shows the asset-weighted relative mispricing for all U.S. based equity index ETFs. The relative mispricing is defined as the difference between the ETF closing market price and its NAV at the end of the day and expressed in percent of the fund NAV. The estimates are based on the daily sample of CRSP index ETFs. Funds are sorted into asset classes based on the Morningstar classification.

Figure A.6: Distribution of Relative Mispricing for Equity Index ETFs



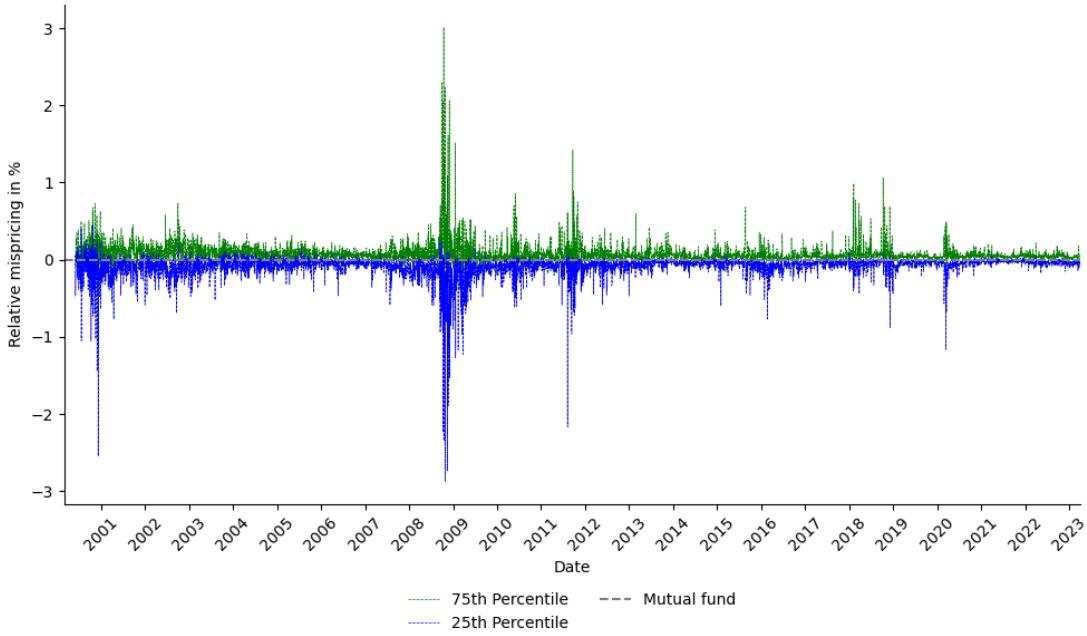
Note: This figure shows the cross-sectional distribution of the relative mispricing for the sample of U.S. based equity index ETFs underlying figure A.5.

Figure A.7: Asset-Weighted Relative Mispricing for Domestic Equity Index ETFs



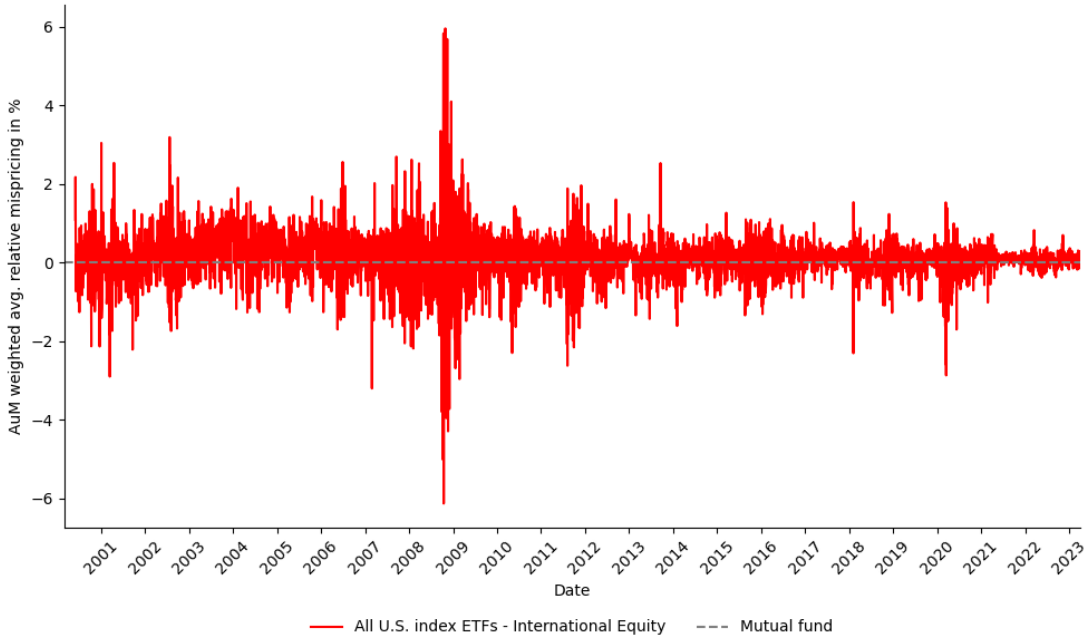
Note: This figure shows the asset-weighted relative mispricing for all U.S. based domestic equity index ETFs. The relative mispricing is defined as the difference between the ETF market price and its NAV in percent of the fund NAV. The estimates are based on the daily sample of CRSP index ETFs. Funds are classified into asset classes based on data from Morningstar.

Figure A.8: Distribution of Relative Mispricing for Domestic Equity Index ETFs



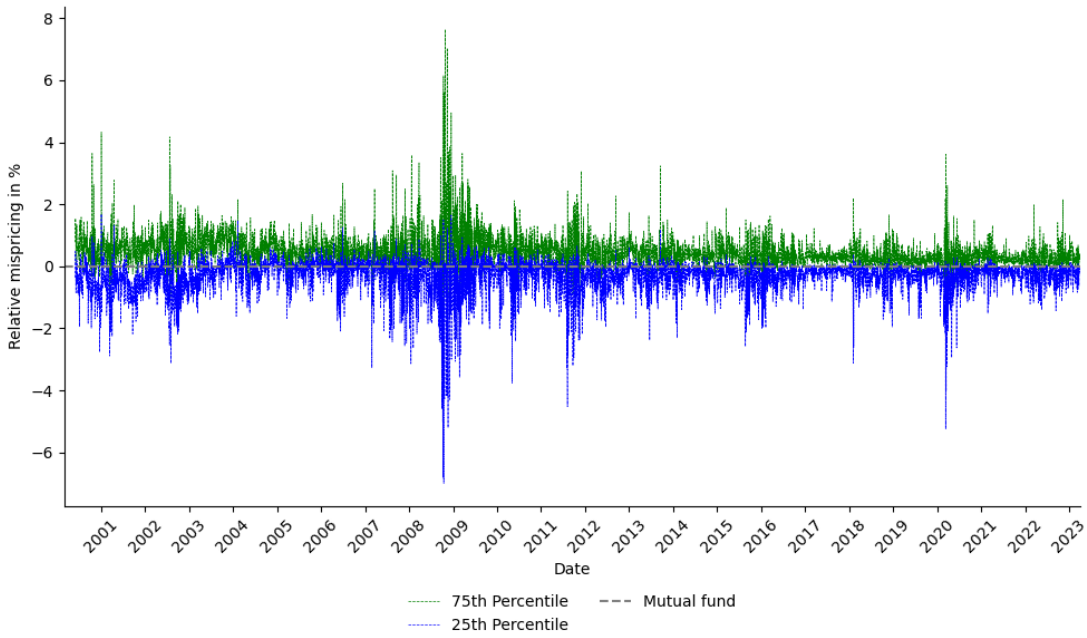
Note: This figure shows the cross-sectional distribution of the relative mispricing for the sample of U.S. based domestic equity index ETFs underlying figure A.7.

Figure A.9: Asset-Weighted Relative Mispricing for International Equity Index ETFs



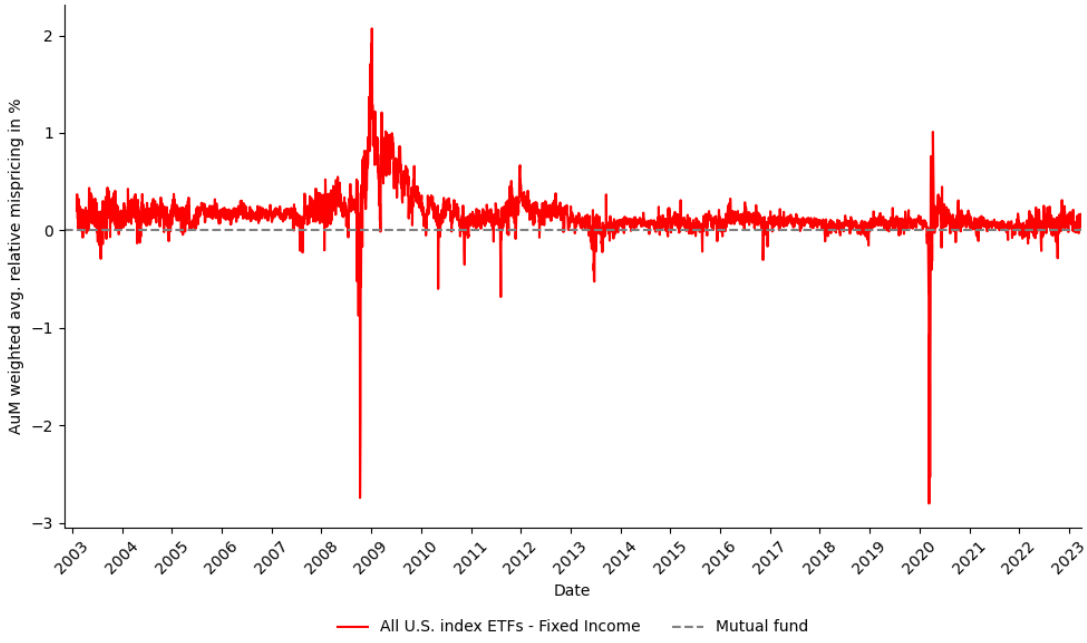
Note: This figure shows the asset-weighted relative mispricing for all U.S. based international equity index ETFs. The relative mispricing is defined as the difference between the ETF market price and its NAV in percent of the fund NAV. The estimates are based on the daily sample of CRSP index ETFs. Funds are classified into asset classes based on data from Morningstar.

Figure A.10: Distribution of Relative Mispricing for International Equity Index ETFs



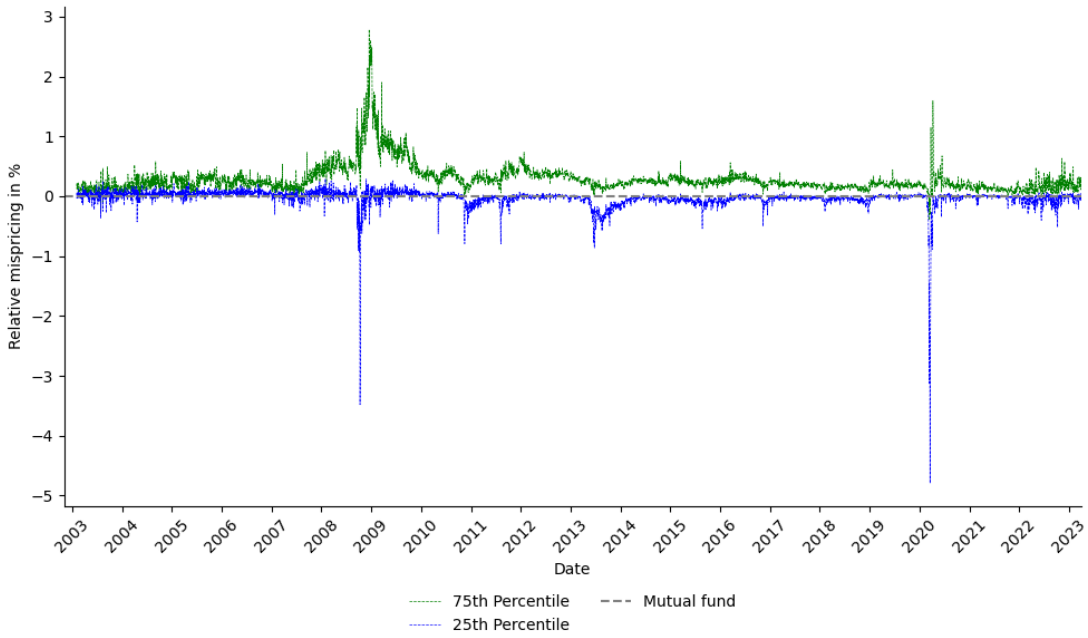
Note: This figure shows the cross-sectional distribution of the relative mispricing for the sample of U.S. based international index equity ETFs underlying figure A.9.

Figure A.11: Asset-Weighted Relative Mispricing for Fixed Income ETFs



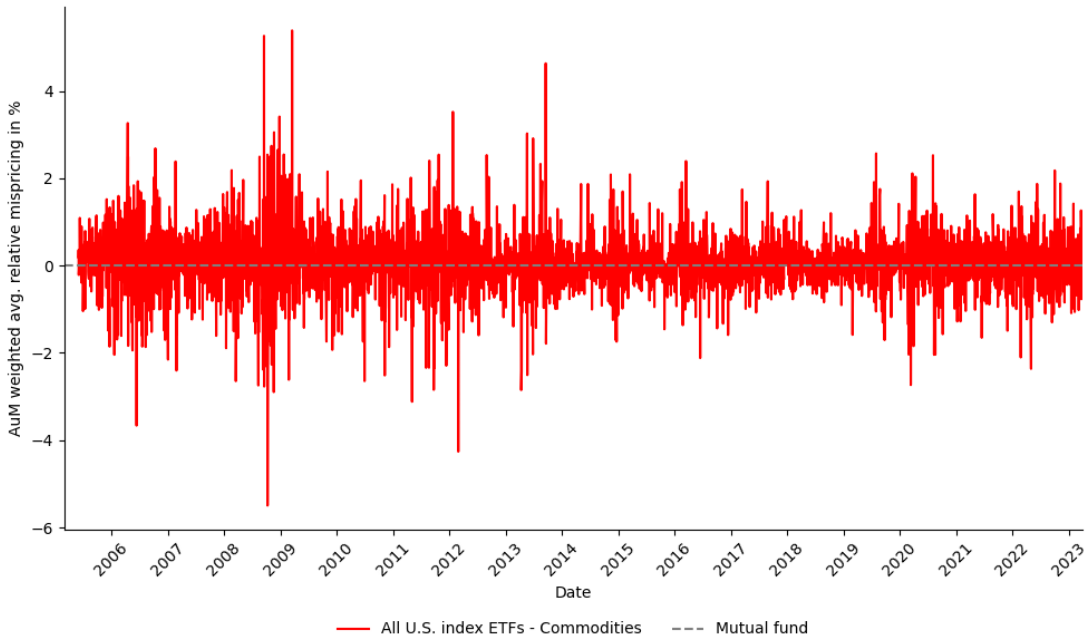
Note: This figure shows the asset-weighted relative mispricing for all U.S. based fixed income index ETFs. The relative mispricing is defined as the difference between the ETF market price and its NAV in percent of the fund NAV. The estimates are based on the daily sample of CRSP index ETFs. Funds are classified into asset classes based on data from Morningstar.

Figure A.12: Distribution of Relative Mispricing for Fixed Income Index ETFs



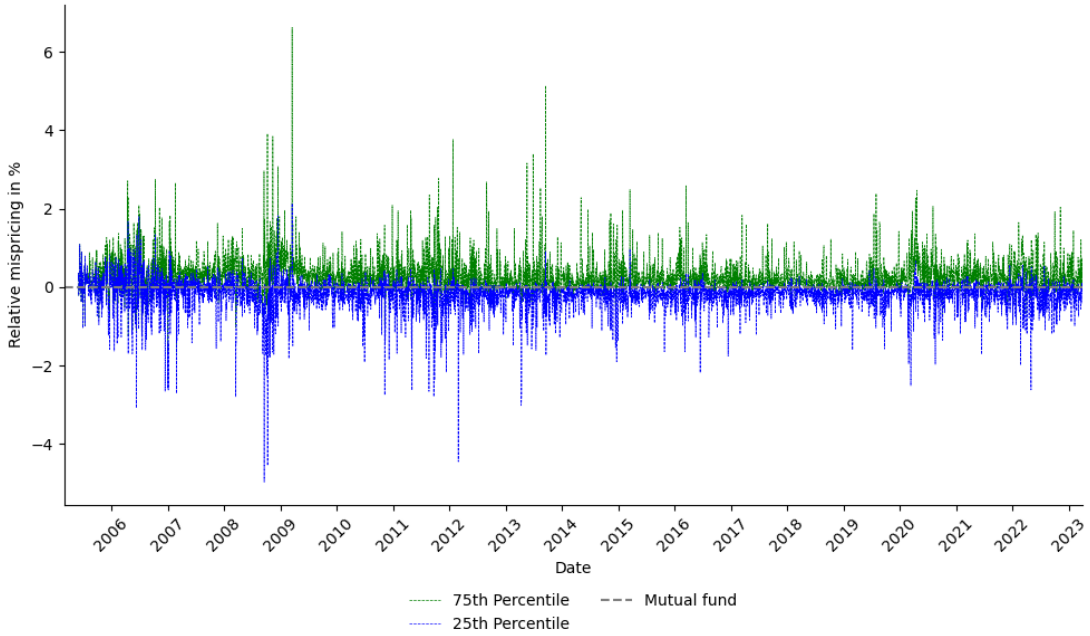
Note: This figure shows the cross-sectional distribution of the relative mispricing for the sample of U.S. based fixed income index ETFs underlying figure A.11.

Figure A.13: Asset-Weighted Relative Mispricing for Commodity ETFs



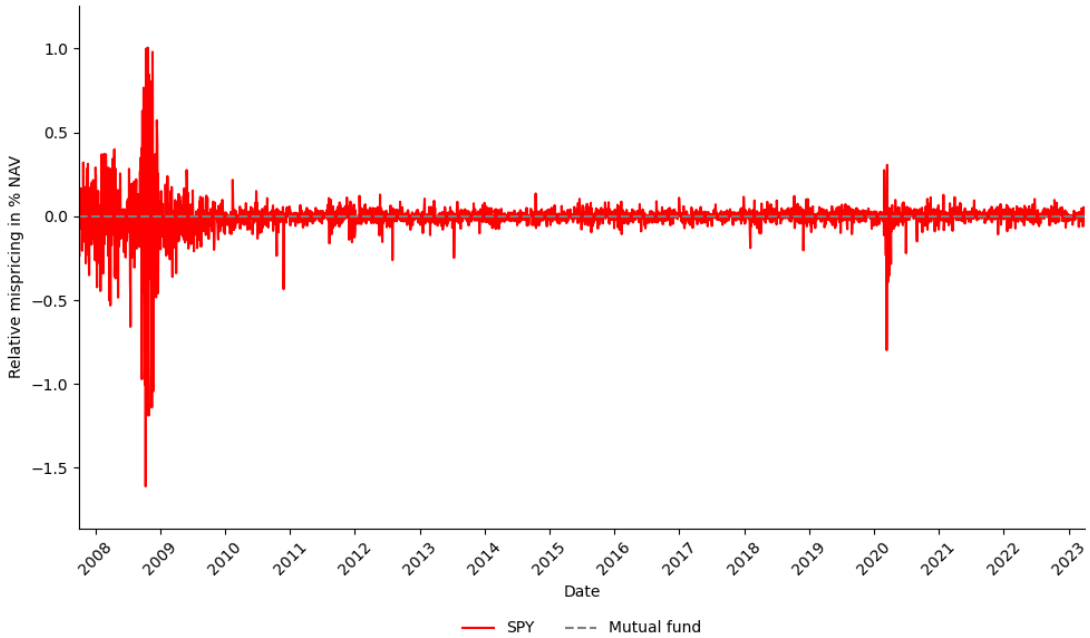
Note: This figure shows the asset-weighted relative mispricing for all U.S. based commodity index ETFs. The relative mispricing is defined as the difference between the ETF market price and its NAV in percent of the fund NAV. The estimates are based on the daily sample of CRSP index ETFs. Funds are classified into asset classes based on data from Morningstar.

Figure A.14: Distribution of Relative Mispricing for Commodity Index ETFs



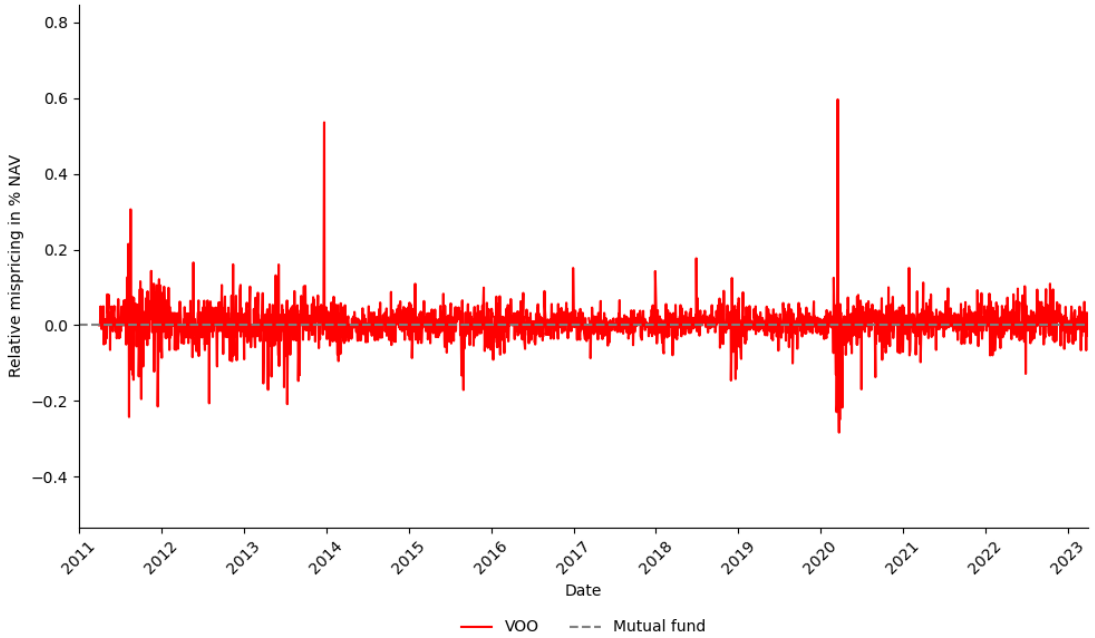
Note: This figure shows the cross-sectional distribution of the relative mispricing for the sample of U.S. based commodity index ETFs underlying figure A.11.

Figure A.15: Relative Mispricing for SPY ETF



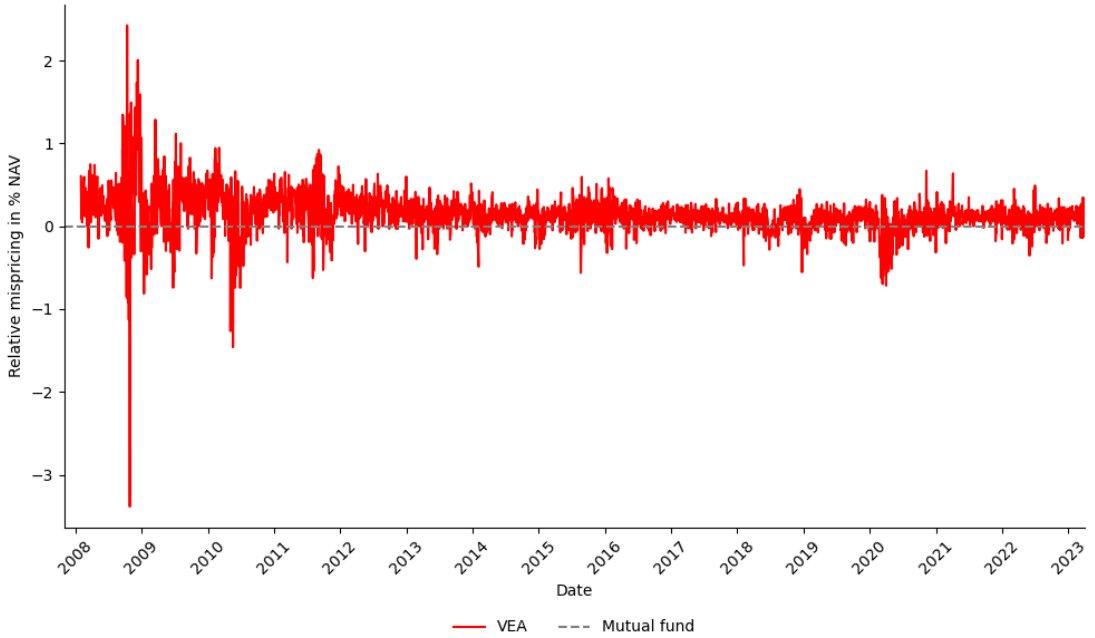
Note: This figure shows the relative mispricing for the SPDR S&P 500 ETF Trust, the largest S&P 500 ETF with ticker SPY. The relative mispricing is defined as above. The sample mean (median) relative mispricing is −0.29 bps (0.14 bps) with a standard deviation of 10.04 bps.

Figure A.16: Relative Mispricing for VOO ETF



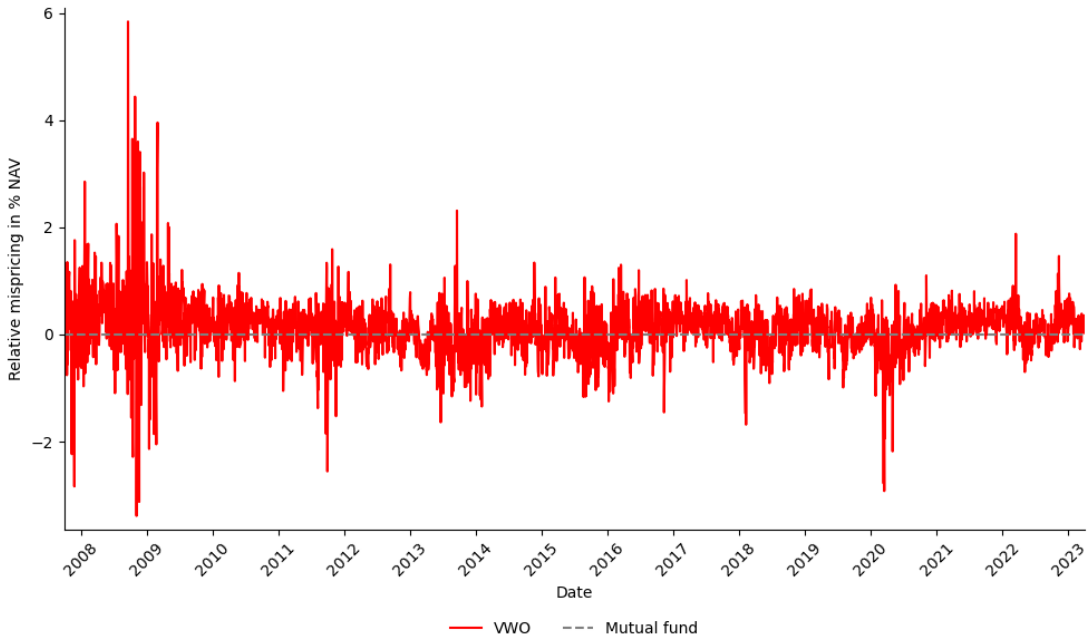
Note: This figure shows the relative mispricing for the Vanguard S&P 500 ETF, the third largest S&P 500 ETF with ticker VOO. The relative mispricing is defined as above. The sample mean (median) relative mispricing is 0.5 bps (0.54 bps) with a standard deviation of 4.09 bps.

Figure A.17: Relative Mispricing for VEA ETF



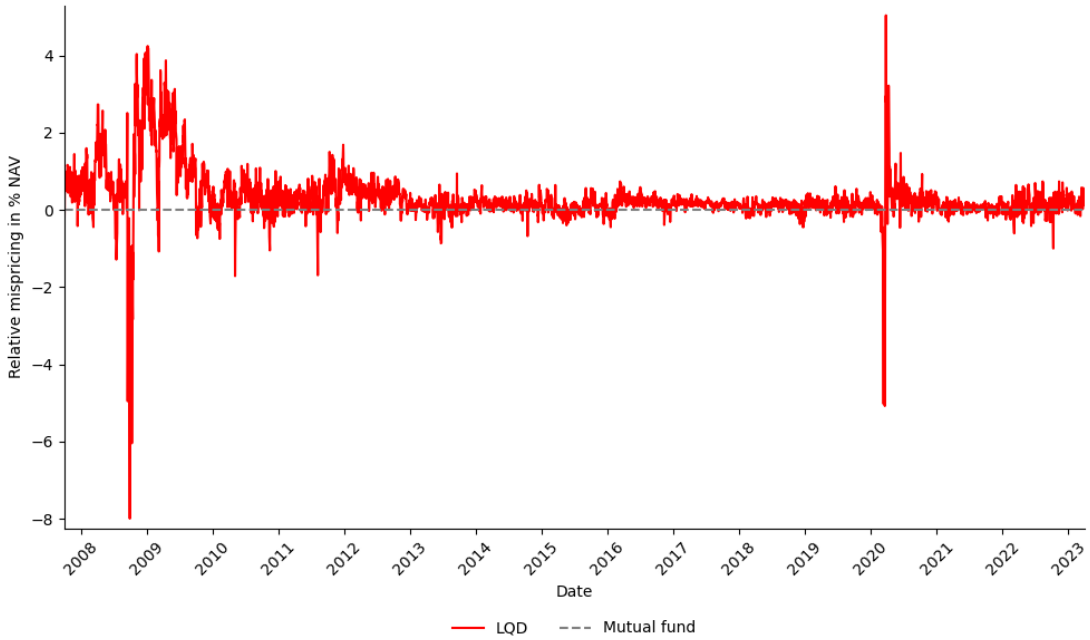
Note: This figure shows the relative mispricing for the Vanguard FTSE Developed Markets ETF, the largest international equity ETF with ticker VEA. The relative mispricing is defined as above. The sample mean (median) relative mispricing is 14.67 bps (13.35 bps) with a standard deviation of 23.19 bps.

Figure A.18: Relative Mispricing for VWO ETF



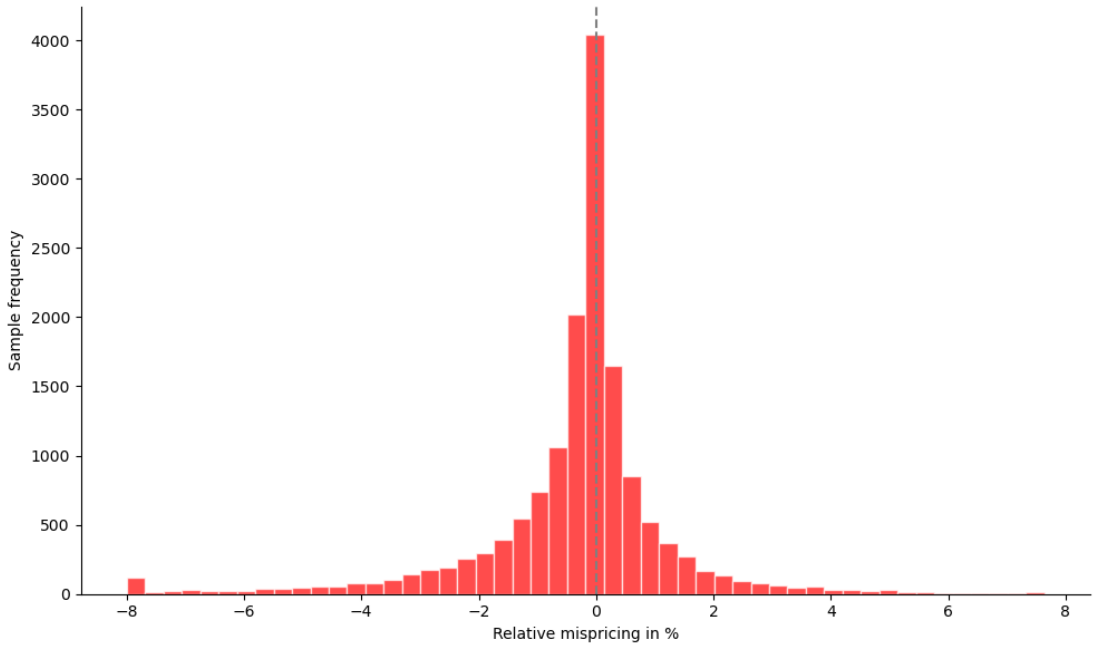
Note: This figure shows the relative mispricing for the Vanguard Emerging Markets Stock Index Fund ETF, the largest emerging markets equity ETF with ticker VWO. The relative mispricing is defined as above. The sample mean (median) relative mispricing is 11.19 bps (14.01 bps) with a standard deviation of 49.42 bps.

Figure A.19: Relative Mispricing for LQD ETF



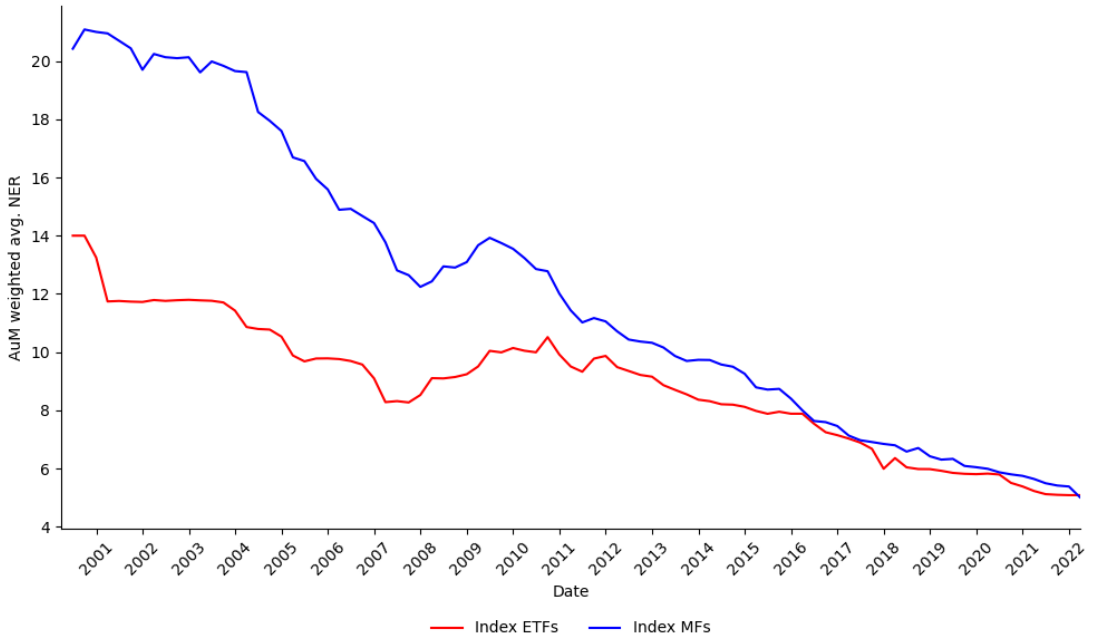
Note: This figure shows the relative mispricing for the iShares iBoxx $ Investment Grade Corporate Bond ETF, one of the most liquid investment grade corporate bond ETFs with ticker LQD. The relative mispricing is defined as above. The sample mean (median) relative mispricing is 30.42 bps (17.34 bps) with a standard deviation of 67.142 bps.

Figure A.20: Distribution of Relative Mispricing in U.S. Index ETFs in March 2020



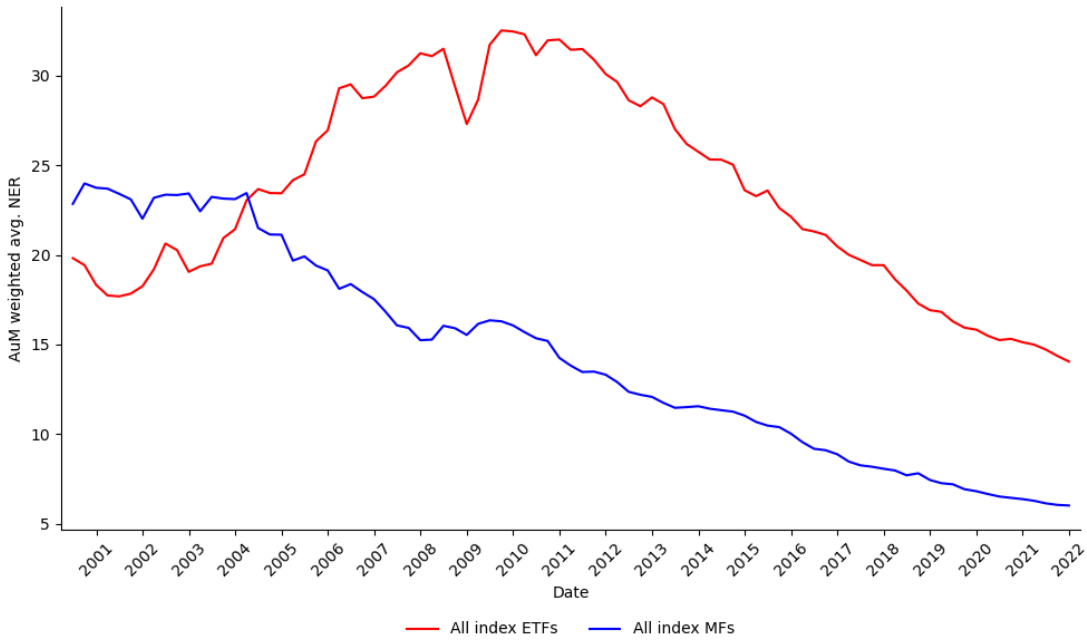
Note: This figure shows the distribution of relative mispricing for the sample of all U.S. based index ETFs during the March 2020 market turmoil. The mean (median) mispricing amounted to −32.5 bps (−8.3 bps), with a standard deviation of 164.8 bps. The relative mispricing is defined as above.

Figure A.21: Asset-Weighted Net Expense Ratio in Top 5 Benchmark Segments



Note: This figure shows the asset-weighted net expense ratio for U.S. based index ETFs and open-end mutual funds tracking one of the top five benchmark segments in basis points. The top five benchmark indices are based on the average AUM of funds tracking a given benchmark index over the sample between Q2 2000 and Q1 2023 and include the S&P 500, CRSP US Total Market, Bloomberg US Aggregate Float Adjusted, FTSE Global All Cap ex US, and Bloomberg Global Aggregate ex-USD index. The estimates are based on the quarterly sample of CRSP index funds with benchmark index classifications from Morningstar.

Figure A.22: Asset-Weighted Net Expense Ratio in All Benchmark Segments



Note: This figure shows the asset-weighted net expense ratio for all U.S. based index ETFs and open-end mutual funds in basis points. The estimates are based on the quarterly sample of CRSP index funds with benchmark index classifications from Morningstar.

B Institutional background

This section summarizes the key institutional details underlying ETFs and open-end mutual funds. I focus on the structure of U.S. markets. Apart from differences in fund taxation and the reporting of fund flows from intermediaries to fund sponsors, ETFs and MFs generally function similarly in international markets. For a more detailed summary of ETFs I refer to Ben-David, Franzoni, and Moussawi (2017) and Lettau and Madhavan (2018).

The first U.S. mutual fund was launched in 1924 (Federal Reserve 2000). For most of history, mutual funds have been the only way for retail investors as well as many institutional investors to obtain cheap portfolio diversification and access less liquid market segments.[21] MFs were a revolution because they brought the average person into the stock market. In contrast, ETFs were originally established in 1993 by stock exchanges and targeted at traders of futures contracts. ETFs simply represented a new way for investors to trade bundles of stocks. Initially, they were not intended to directly compete with MFs as a long-term investment vehicle for the average investor. This is consistent with the idea of ETFs being tailored towards short-term traders who require intraday liquidity, a view that remains widespread in the academic literature. Yet, ever since the financial crisis in 2008, driven by the shift from

[21] Another instrument to obtain index exposure without directly holding the index constituent securities are futures contracts. Yet, given the additional requirements (e.g., margin account) and risks associated with derivatives trading, index futures do not constitute a significant alternative to ETFs and MFs in the retail investing space.

active to passive investing, ETFs have become more than just a vehicle for high turnover trading strategies. ETFs now constitute a popular alternative investment vehicle for many different types of investors. In response to investors' increasing demand for ETFs, some asset managers have already converted existing MFs into ETFs.

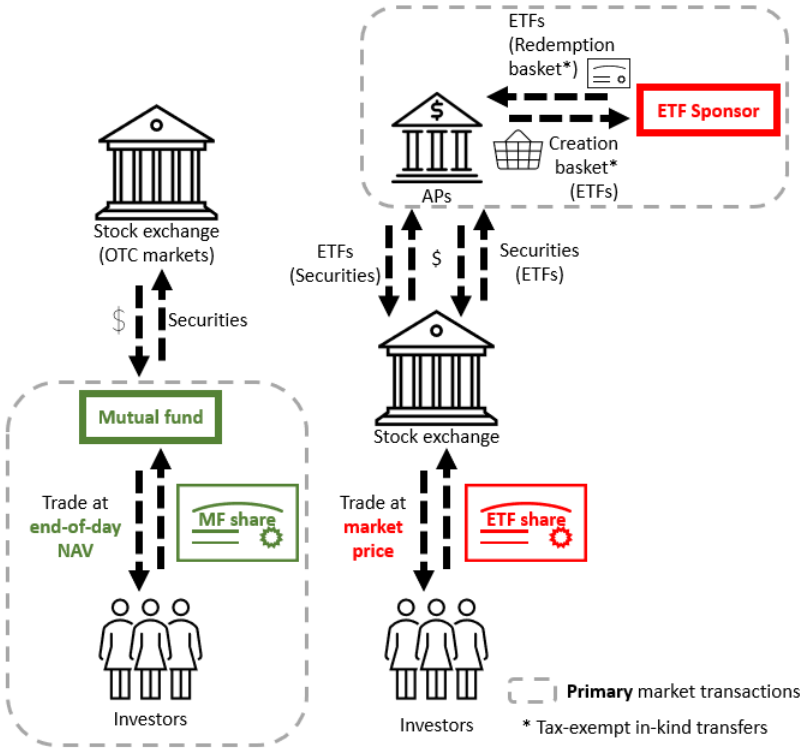
Figure B.8 illustrates how MFs and ETFs are priced and traded in financial markets. MF shares are purchased or redeemed directly from the fund sponsor at the fund NAV. Unlike ETFs, all MF trades submitted during the trading day are executed at the same price, the end-of-day fund NAV. In the U.S., orders for MFs must usually be submitted by 4pm ET to be executed at the same-day fund NAV. Any orders submitted thereafter will be executed at the next available NAV, so on the next trading day. It is noteworthy that MF transactions in the U.S. generally settle $T+1$ or $T+2$, depending on the fund type. While the settlement period of equities, including ETFs, is currently $T+2$, the U.S. is also moving towards $T+1$ settlement in these markets as of 28 May 2024.[22] By construction, MFs have zero relative mispricing. The end-of-day MF NAV is determined based on the closing prices of the fund's portfolio securities (or their estimates). Therefore, MF prices may not fully incorporate the price impact and trading costs that ensue when funds liquidate assets to satisfy investor redemptions on the next trading day. The MF can satisfy net investor redemptions by temporarily depleting cash buffers, if available, but will eventually need to liquidate asset holdings when outflows are large. MFs can generally not satisfy redemptions using in-kind transfers (RIK) of security baskets. RIKs are only feasible for large investors, but MFs rarely make use of the option to meet redemptions in-kind even if they are legally permitted to do so.[23] Thus, by construction, net MF redemption are directly linked to transactions in portfolio securities and costly for the fund. Flow-induced transaction costs for MFs include commissions, bid-ask spreads, price impact and taxes on capital gains distributions. When capital gains are realized as a result of security sales after redemptions, these taxes are borne by the remaining fund investors. This implies that MF investors may be subject to an early realization of capital gains taxes even if they remain invested in the fund. Overall, because of how MF shares are priced, transaction costs are caused by exiting investors but borne by the remaining MF shareholders.

ETFs are traded intraday in secondary markets at the prevailing market price. Secondary market transactions of ETF shares occur between investors, potentially via a market maker. Investors do not directly trade with the ETF sponsor. Therefore, ETF trades are not directly linked to transactions in portfolio securities. Instead, the ETF's market price is indirectly linked to its NAV and asset markets via the law of one price and the arbitrage trades con-

[22] See SEC Release Nos. 34-96930. These settlement dates refer to the time and when the ownership in a financial instrument is transferred.

[23] Based on data from U.S. funds' shareholder reports, Agarwal, Ren, Shen, and Zhao (2022) find that only 13.1% of the funds which reserve the right to execute redemptions in kind actually engaged in in-kind redemptions at least once during a sample from 1997 to 2017.

Figure B.8: Trading of ETFs and open-end Mutual Funds



ducted by APs.[24] APs are financial institutions, usually large broker-dealers, with the right, but not the obligation, to create and redeem ETF shares outright. For example, when the ETF price exceeds the fund NAV, they can deliver a basket of securities, the creation basket, to the fund sponsor in exchange for new ETF shares, keeping the price difference as an arbitrage profit. Generally, creation and redemption baskets resemble the portfolio securities held by the ETF. Through this process, APs increase the number of ETF shares outstanding. Importantly, creations and redemptions of ETF shares normally do not involve cash but occur in-kind. As a special case, in the U.S. such in-kind transfers of securities are also tax-exempt.[25] This mechanism turns APs into the central suppliers of liquidity in ETF markets. At the same time, the dependence on AP arbitrage renders ETFs vulnerable to shocks to financial intermediaries' balance sheet capacity. When APs temporarily retreat from ETF creations or redemptions, ETF prices can start deviating substantially from the fund NAV, leading to relative mispricing as shown in figures A.5 - A.14. This potential discrepancy between the price at which investors can liquidate ETF shares at short notice and the fund

[24]The described mechanism refers to ETFs that physically replicate their benchmark index by holding the underlying securities. There also exist synthetic ETFs that track their benchmark index using derivatives, such as swaps. Yet, the large majority of ETFs pursue the physical index replication process which this paper builds upon.

[25]Due to their in-kind creation and redemption mechanism, capital gains taxes on ETFs are deferred until shares are sold by the investor. As a result, ETFs rarely distribute capital gains allowing investors to defer capital gains taxes until they liquidate their shareholdings. U.S.-based mutual funds must distribute any capital gains to shareholders at least once a year. These distributions are taxable if MFs shares are held within taxable accounts.

NAV constitutes the key friction in ETF markets.

C Proofs

Proof of Lemma 1. The only way for a non-zero tracking difference to occur in ETFs is for creation and redemption baskets to diverge from the underlying benchmark index. This is ruled out by assumption 1. Hence, one ETF share is always equivalent to one unit of the composite security, implying zero tracking difference.

Proof of Lemma 2. By definition $\Delta_2^M \equiv P_2^j - NAV_2^M$. The terminal index price is given exogenously by $P_2^j = x_j$. The MF NAV at the end of any given trading day is defined by the accounting identity $NAV_t^{M,j} \equiv \frac{X_t^{M,j} P_t^j}{\kappa_t^{M,j}}$, where $X_t^{M,j}$ is the number of index shares held by the fund, P_t^j is the market price per unit of index share and $\kappa_t^{M,j}$ is the remaining number of MF shares outstanding.

In the model, the number of index shares held by the MF at $t = 1$ and $t = 2$ are given by

$$X_1^{M,j} = \kappa_0^{M,j}, \tag{41}$$

$$X_2^{M,j} = \kappa_0^{M,j} - \underbrace{\frac{\psi x_j \Delta\kappa_1^{M,j}}{P_{1^+}^{j,M}}}_{\substack{\text{Flow-induced} \\ \text{index sales}}}, \tag{42}$$

where $\Theta_{1^+}^{j,M} = \frac{\psi x_j \Delta\kappa_1^{M,j}}{P_{1^+}^{j,M}}$ is the number of index shares sold by the MF at $t = 1^+$ to meet net fund redemptions in the interim period as derived in equation 24.

It follows

$$\begin{aligned}
\Delta_2^{M,j} &= P_2^j - NAV_2^{M,j} \\
&= P_2^j - \frac{P_2^j X_2^{M,j}}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} \\
&= P_2^j - \frac{P_2^j(\kappa_0^{M,j} - \Theta_{1^+}^{j,M})}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}.
\end{aligned} \tag{43}$$

Substituting the expression for $\Theta_{1^+}^{j,M}$ from equation 24,

$$
\begin{aligned}
\Delta_2^{M,j} &= P_2^j - \underbrace{\frac{P_2^j(\kappa_0^{M,j} - \frac{\Delta\kappa_1^{M,j}\psi x_j}{P_{1^+}^{j,M}})}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}}_{NAV_2^M} \\
&= P_2^j - P_2^j\left(\frac{\kappa_0^{M,j}}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} - \frac{\left(\frac{\Delta\kappa_1^{M,j}\psi x_j}{P_{1^+}^{j,M}}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}\right) \\
&= P_2^j - P_2^j\left(\frac{(\kappa_0^{M,j} - \Delta\kappa_1^{M,j} + \Delta\kappa_1^{M,j})}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} - \frac{\left(\frac{\Delta\kappa_1^{M,j}\psi x_j}{P_{1^+}^{j,M}}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}\right) \\
&= P_2^j - P_2^j\left(1 - \frac{\left(\frac{\Delta\kappa_1^{M,j}\psi x_j}{P_{1^+}^{j,M}} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}\right) \qquad (44)
\end{aligned}
$$

Finally,

$$
\Delta_2^{M,j} = \frac{P_2^j\left(\frac{\Delta\kappa_1^{M,j}\psi x_j}{P_{1^+}^{j,M}} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} \qquad (45)
$$

The MF tracking difference represents the absolute cost of the share dilution per unit of MF share experienced by remaining MF shareholders. It can expressed as a fraction of the terminal index value (the numeraire):

$$
\tilde{\Delta}_2^{M,j} \equiv \frac{\Delta_2^{M,j}}{P_2^j} = \frac{\Delta\kappa_1^{M,j}\psi x_j (P_{1^+}^{j,M})^{-1} - \Delta\kappa_1^{M,j}}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}}, \qquad (46)
$$

The relative MF share dilution per unit of MF shares in 46 can further be decomposed into three components:

$$
\tilde{\Delta}_2^M = \underbrace{\frac{1}{\kappa_1^{M,j}}}_{\substack{\text{Remaining} \\ \text{MF shares} \\ \text{outstanding}}} \left(\underbrace{\Delta\kappa_1^{M,j}}_{\substack{\text{\# MF} \\ \text{shares} \\ \text{redeemed} \\ \text{early}}} \left(\underbrace{\psi x_j (P_{1^+}^{j,M})^{-1} - 1}_{\substack{\text{Excess \# index} \\ \text{shares liquidated} \\ \text{to satisfy early} \\ \text{redemptions}}} \right) \right). \qquad (47)
$$

Proof of Corollary 1. First, from equation 45 it directly follows that $\Delta_2^{M,j} = 0$ if net mutual fund redemptions at $t = 1$ are zero, $\Delta\kappa_1^{M,j} = 0$.

Second, if the mutual fund NAV at $t = 1$ is perfectly forward looking and equal to the index price at which the MF trades in index markets, $\psi_j x_j = P_{1^+}^{j,M}$, $\Delta_2^{M,j} = 0$ because

$$\Delta_2^{M,j} = \frac{P_2^j\left(\frac{\Delta\kappa_1^{M,j}\psi_j x_j}{P_{1^+}^{j,M}} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} = \frac{P_2^j\left(\Delta\kappa_1^{M,j} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} = 0$$

Third, if markets are perfectly liquid and funds have no price impact when trading index shares, $\psi_j = 1$ and $c_j = 0$. As a result, $NAV_1^{M,j} = x_j$ and $P_{1^+}^{j,M} = x_j$. It follows

$$\Delta_2^{M,j} = \frac{P_2^j\left(\frac{\Delta\kappa_1^{M,j}\psi_j x_j}{P_{1^+}^{j,M}} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} = \frac{P_2^j\left(\frac{\Delta\kappa_1^{M,j} x_j}{x_j} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} = \frac{P_2^j\left(\Delta\kappa_1^{M,j} - \Delta\kappa_1^{M,j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} = 0$$

Finally, if $\Delta\kappa_1^{M,j} \neq 0$ and $NAV_1^{M,j} \neq P_{1^+}^{j,M}$, then $\Delta_2^{M,j} \neq 0$. This follows directly from equation 47 combined with the facts that $x_j > 0$ and $\kappa_1^{M,j} = \eta + \kappa_0^{M,j} - \Delta\kappa_1^{M,j} > 0$.

Proof of Corollary 2. From proof 1 it follows that $\Delta_2^{M,j} = 0$ when at least one of the conditions in corollary 1 is satisfied. As a result,

$$P_2^{M,j} = P_2^j - \Delta_2^{M,j} = P_2^j. \tag{48}$$

Since $P_2^{E,j} = P_2^j$, $P_2^{M,j} = P_2^{E,j} = P_2^j$.

Proof of Lemma 3. I start by proving the results (i) and (iii): Conditional on receiving no liquidity shock, ETF (MF) investors are ex-post identical. This is because at $t = 0$ investors only had the choice between investing their entire, identical endowment into MFs or ETFs. Investors did not have the option to invest parts of their endowment in the risk-free asset. Due to their risk-neutrality, investors never choose to mix between the ETF and MF at $t = 0$. All ETF (MF) investors at $t = 0$ with $i = l$ at $t = 1$, that is patient ETF (MF) investors within a given index segment j, have the same initial allocation $\theta_0^{i,E,j} = 1$ ($\theta_0^{i,M,j} = 1$). This allocation implies that all patient ETF (MF) investors' budget constraints at $t = 1$ must be identical. In addition, by assumption, all investors have identical preferences over terminal wealth. Finally, in equilibrium, all MF investors share the same belief regarding other investors' redemption decisions. Thus, all remaining patient ETF (MF) investors must have the same optimal investment policy at $t = 1$, that is

$$\theta_1^{i,E,j^*} = \theta_1^{E,j^*} \;\forall\; i \text{ with } \theta_0^{i,E,j} = 1 \text{ and } i = l \tag{49}$$

$$\theta_1^{i,M,j^*} = \theta_1^{M,j^*} \;\forall\; i \text{ with } \theta_0^{i,M,j} = 1 \text{ and } i = l \tag{50}$$

Result (ii) and (iv) follow the above argument: Conditional on receiving a liquidity shock, ETF (MF) investors are ex-post identical. All ETF (MF) investors at $t = 0$, have the same initial allocation $\theta_0^{i,E,j} = 1$ ($\theta_0^{i,M,j} = 1$). Hence, all ETF (MF) investors with $i = e$ at $t = 1$

face the same budget constraint. When liquidating their entire portfolio upon receiving a liquidity shock at $t = 1$, all impatient ETF (MF) receive identical payoffs. Thus,

$$\theta_1^{i,E^*} = 0 \ \forall \ i \text{ with } \theta_0^{i,E} = 1 \text{ and } i = e \tag{51}$$

$$\theta_1^{i,M^*} = 0 \ \forall \ i \text{ with } \theta_0^{i,M} = 1 \text{ and } i = e \tag{52}$$

Their terminal wealth is given by

$$W_1^{i*} = P_1^{E^*} \ \forall \ i \text{ with } \theta_0^{i,E} = 1 \text{ and } i = e \tag{53}$$

$$W_1^{i*} = NAV_1^{M^*} \ \forall \ i \text{ with } \theta_0^{i,M} = 1 \text{ and } i = e \tag{54}$$

Proof of Proposition 1. Since investors are risk-neutral, patient ETF investors do not liquidate any of their fund shares at $t = 1$, $\theta_1^{i,E,j} = 1 \ \forall i = l$, whenever the expected ETF payoff at $t = 2$ exceeds the ETF price at $t = 1$. Formally, when

$$P_1^{E,j} \leq E[P_2^{E,j}] \tag{55}$$

, where the weak nature of the inequality follows directly from assumption 3.

Using assumption 2 and the expression for the market-clearing ETF price at $t = 1$ from equation 22, equation 55 reduces to

$$0 \leq \Delta\kappa_1^{E,j}(c_j + \phi_j). \tag{56}$$

The market maker's inventory cost parameter as well as AP's balance sheet capacity constraint parameter are both generally strictly positive, $c_j + \phi_j > 0$. Besides, the model only features negative liquidity shocks. There are no positive liquidity (savings) shocks. There are no dividends or non-financial income sources. As a result, $\Delta\kappa_1^{E,j} \geq 0$, there can only be outflows. Hence, equation 56 is always satisfied and patient ETF investors always remain invested until the terminal period.

The special case in which equation 56 is satisfied with equality occurs when $\Delta\kappa_1^{E,j}$ is zero or when c_j and ϕ_j are both zero. $\Delta\kappa_1^{E,j} = 0$ holds when none of the investors who invested in ETFs at $t = 0$ receives a liquidity shock at $t = 1$. $c_j = 0$ and $\phi_j = 0$ hold in the stylist scenario case in which index markets are very liquid, such that the market maker can absorb any index supply without price impact and the AP faces no balance sheet capacity constraints.

Proof of Proposition 2. The payoffs of the ETF and MF tracking index j at different horizons are given by

$$P_1^{E,j} = x_j - \Delta\kappa_1^{E,j}(c_j + \phi_j) \tag{57}$$
$$P_2^{E,j} = x_j \tag{58}$$
$$P_1^{M,j} = \psi x_j \tag{59}$$
$$P_2^{M,j} = x_j - \Delta_2^{M,j} \tag{60}$$

(i) When $c_j = \psi_j = \phi_j = 0$, the ETF and MF price at $t = 1$ reduce to x_j, the fundamental index value. Besides, from corollaries 1 and 2 follows that $\Delta_2^{M,j} = 0$. Hence, the ETF and MF payoff at $t = 2$ are identical as well.

(ii) Following corollary 1, if the MF NAV at $t = 1$ is perfectly forward looking, $P_1^{M,j} = P_{1^+}^j$, the terminal MF tracking difference is zero, $\Delta_2^{M,j} = 0$. Thus, $P_2^{E,j} = P_2^{M,j} = x_j$.

In the absence of AP balance sheet capacity constraints, $\phi_j = 0$, the ETF price at $t = 1$ becomes $P_1^{E,j} = x_j - \Delta\kappa_1^{E,j} c_j$. By assumption, in this case $c_j > 0$. Defining $\Delta\kappa_1^{E,j} = \Delta\kappa_1^{M,j} = \Delta\kappa_1^j$, the fund payoffs at $t = 1$ simplify to

$$P_1^{E,j} = x_j - \Delta\kappa_1^j c_j \tag{61}$$
$$P_1^{M,j} = P_{1^+}^j = x_j - \Delta\kappa_1^j c_j \tag{62}$$

Equation 62 follows directly from market clearing between the MF and index market maker at $t = 1^+$ according to equation 25 for the case that $P_1^{M,j} = P_{1^+}^j$.

Proof of Lemma 4.

Since investors are risk-neutral, patient MF investors do not liquidate any of their fund shares at $t = 1$, $\theta_1^{i,M,j} = 1 \ \forall i = l$, whenever the expected MF payoff at $t = 2$ exceeds the MF price at $t = 1$. Formally, when

$$P_1^{M,j} \leq E[P_2^{M,j} | x_j]. \tag{63}$$

Using the definition of the MF NAV at $t = 1$ and the results from equations 15 and 17, 63 reduces to

$$\psi x_j \leq E\left[P_2^j - \frac{P_2^j \left(\frac{\Delta\kappa_1^{M,j}(NAV_1^{M,j} - P_{1^+}^j)}{P_{1^+}^j} \right)}{\kappa_1^{M,j}} \middle| x_j \right]. \tag{64}$$

Simplifying and using $x_j > 0$,

$$\psi \leq 1 - \frac{\left(\frac{E[\Delta\kappa_1^{M,j}|x_j](\psi x_j - P_{1^+}^j)}{P_{1^+}^j}\right)}{E[\kappa_1^{M,j}|x_j]}. \tag{65}$$

If index markets are frictionless, $c_j = 0$, there is no price impact as the index market maker absorbs any supply of index shares at $P_{1^+}^j = x_j$. It follows:

$$\psi \leq 1 - \frac{\left(\frac{E[\Delta\kappa_1^{M,j}|x_j](\psi x_j - x_j)}{x_j}\right)}{E[\kappa_1^{M,j}|x_j]}. \tag{66}$$

This in turn reduces to

$$\left(1 + \frac{E[\Delta\kappa_1^{M,j}|x_j]}{E[\kappa_1^{M,j}|x_j]}\right)\psi \leq 1 + \frac{E[\Delta\kappa_1^{M,j}|x_j]}{E[\kappa_1^{M,j}|x_j]}. \tag{67}$$

Since there can only be outflows or zero fund flows at $t = 1$, $\Delta\kappa_1^{M,j} \geq 0$. In the presence of sleepy investors, $\eta > 0$, the number of remaining MF investors is strictly positive, $\kappa_1^{M,j} > 0$. Hence, $1 + \frac{E[\Delta\kappa_1^{M,j}|x_j]}{E[\kappa_1^{M,j}|x_j]}$ is strictly positive. Because $0 < \psi < 1$, equation 67 always holds and, irrespective of other MF investors' redemptions, $\Delta\kappa_1^{M,j}$, patient MF investors would never liquidate any shares early if $c_j = 0$.

Proof or Proposition 3. Market clearing in index markets between index market makers and MFs, $\Theta_{1^+}^{M,j} = \Theta_{1^+}^{D,j}$, implies

$$\frac{E[P_2^j|x_j] - P_{1^+}^{j,M}}{c_j} = \frac{\Delta\kappa_1^{M,j} P_1^M}{P_{1^+}^{j,M}} \tag{68}$$

$$\iff \frac{x_j - P_{1^+}^{j,M}}{c_j} = \frac{\Delta\kappa_1^{M,j}\psi_j x_j}{P_{1^+}^{j,M}}$$

$$\iff 0 = P_{1^+}^{j,M\,2} - P_{1^+}^{j,M} x_j + c_j\psi_j x_j \Delta\kappa_1^{M,j} \tag{69}$$

.

Using assumption 4, it follows $P_{1^+}^{j,M*} = \frac{1}{2}(x_j + \sqrt{x_j^2 - 4c_j\psi_j x_j \Delta\kappa_1^{M,j}})$.

Proof of Corollary 3. Using $\Delta\kappa_1^{M,j} \leq 1$, there can never be more MF redemptions than initial investments, it follows that the fundamental must satisfy $x_j \geq 4c_j\psi_j$ for the solution to equation 69 to exist. The fundamental must be large relative to the price impact and MF NAV staleness parameter.

Proof of Corollary 4. This result directly follows from equation 68 and proposition 3.

Proof of Corollary 5. These results directly follow from equation 68 and the fact that the inverse net index demand function of the market maker, $P_{1^+}^{D,j}$, and the inverse net index supply by the MF, $P_{1^+}^{M,j}$, are strictly decreasing in the number of index shares. On one side, as the supply of index shares by the MF increases, the market maker's inventory cost increases leading to a drop in its willingness to pay per index share. On the other side, as the index price decreases, the MF has to liquidate more index shares to satisfy its obligations to MF investors who redeemed at $t = 1$.

$$\frac{dP_{1^+}^{D,j}}{d\Theta_{1^+}^{D,j}} = -c_j \tag{70}$$

$$\frac{dP_{1^+}^{M,j}}{d\Theta_{1^+}^{M,j}} = -\frac{\Delta\kappa_1^{M,j} P_1^{M,j}}{(\Theta_{1^+}^{M,j})^2} \tag{71}$$

over the permissible range of $\Theta_{1^+}^{M,j}$.

Proof of Proposition 4. $\underline{x}_j$ must be such that after observing the fundamental $x_j < \underline{x}_j$ patient MF investors always decide to run and redeem their entire portfolio of MF shares early no matter their beliefs of other patient MF shareholders' actions. Patient MF investors will choose to liquidate their shares at $t = 1$ even if they believe only impatient MF sharesholders redeem early. Conversely, for any $x_j > \overline{x}_j$, patient MF investors will never want to redeem any of their shares early. In this region runs never occur in equilibrium.

Lower-dominance region $(0, \underline{x}_j]$**.** Let $\bar{e}^M = E[e^M]$ be the expected mass of impatient (early) MF investors at $t = 1$, with $\bar{e}^M \leq 1$. Using equation 20 and lemma 2, the lower dominance region with respect to the fundamental x_j is then defined by the following condition:

$$0 = E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \underline{x}_j \cup \Delta\kappa_1^{M,j} = \bar{e}^M] \tag{72}$$

Using the restriction $x_j > 0$:

$$0 = 1 - E\left[\frac{\left(\frac{\Delta\kappa_1^{M,j}(P_1^{M,j} - P_{1^+}^j)}{P_{1^+}^j}\right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} \middle| x_j = \underline{x}_j \cup \Delta\kappa_1^{M,j} = \bar{e}^M\right] - \psi_j R^f. \tag{73}$$

Substituting $P_{1^+}^j$ from the index market clearing at $t = 1^+$ following proposition 3:

$$0 = 1 - \frac{\left(\frac{\bar{e}^M(\psi_j \underline{x}_j - \frac{1}{2}(\underline{x}_j + \sqrt{\underline{x}_j^2 - 4c_j\ \psi_j\ \underline{x}_j\ \bar{e}^M}))}{\frac{1}{2}(\underline{x}_j + \sqrt{\underline{x}_j^2 - 4c_j\ \psi_j\ \underline{x}_j\ \bar{e}^M})}\right)}{\kappa_0^{M,j} - \bar{e}^M} - \psi_j R^f. \tag{74}$$

Note that $\kappa_0^{M,j} - \bar{e}^M > 0$ because of the presence of sleepy MF investors. Simplifying:

$$0 = \left((1-\psi_j R^f)(\kappa_0^{M,j} - \bar{e}^M) + \bar{e}^M\right)\left(1 + \sqrt{1 - \frac{4c_j\ \psi_j\ \bar{e}^M}{\underline{x}_j}}\right) - 2\bar{e}^M\psi_j. \tag{75}$$

Solving for $\underline{x}_j$:

$$\left(\frac{2\bar{e}^M\psi_j}{(1-\psi_j R^f)(\kappa_0^{M,j} - \bar{e}^M) + \bar{e}^M} - 1\right)^2 = 1 - \frac{4c_j\ \psi_j\ \bar{e}^M}{\underline{x}_j}. \tag{76}$$

This gives:

$$\underline{x}_j = \frac{4c_j\ \psi_j\ \bar{e}^M}{1 - \left(\frac{2\bar{e}^M\psi_j}{(1-\psi_j R^f)(\kappa_0^{M,j} - \bar{e}^M) + \bar{e}^M} - 1\right)^2}. \tag{77}$$

Using the assumption $R^f = 1$:

$$\underline{x}_j = \frac{4c_j\ \psi_j\ \bar{e}^M(\kappa_0^{M,j} - \psi_j(\kappa_0^{M,j} - \bar{e}^M))^2}{(\kappa_0^{M,j} - \psi_j(\kappa_0^{M,j} - \bar{e}^M))^2 - (2\psi_j\bar{e}^M - (\kappa_0^{M,j} - \psi_j(\kappa_0^{M,j} - \bar{e}^M)))^2} \tag{78}$$

Defining $a = \kappa_0^{M,j} - \psi_j(\kappa_0^{M,j} - \bar{e}^M)$:

$$\underline{x}_j = \frac{4c_j\ \psi_j\ \bar{e}^M a^2}{4a\psi_j\bar{e}^M - 4(\psi_j\bar{e}^M)^2} \tag{79}$$

Note that $a > 0$ because $a = \kappa_0^{M,j} - \psi_j(\kappa_0^{M,j} - \bar{e}^M) = (1-\psi_j)\kappa_0^{M,j} + \psi_j\bar{e}^M)$ with $0 < \psi_j < 1$, $\kappa_0^{M,j} \geq \eta > 0$ and $\bar{e}^M \geq 0$.

For $\bar{e}^M > 0$, equation 79 gives:

$$\underline{x}_j = \frac{c_j(\psi_j\bar{e}^M + (1-\psi_j)\kappa_0^{M,j})^2}{(1-\psi_j)\kappa_0^{M,j}}. \tag{80}$$

In the special case in which $\bar{e}^M = 0$, $\underline{x}_j = 0$ according to equation 77. The lower dominance region is empty. Since there exists a continuum of investors i with $\lambda_i \sim U[0,1]$, the law of large numbers implies that $\bar{e}^M = 0$ only holds in one of two cases: Apart from the sleepy investors, no one invested in MFs at $t = 0$, that is $\kappa_0^{M,j} = \eta$, or only investors characterized by $\lambda_i = 1$ invested in MFs at $t = 0$.

Existence of the lower dominance region. For the lower dominance region to be non-empty it must hold that $\underline{x}_j > 0$. That is

$$0 < \frac{c_j(\psi_j\bar{e}^M + (1-\psi_j)\kappa_0^{M,j})^2}{(1-\psi_j)\kappa_0^{M,j}}. \tag{81}$$

Assuming $\bar{e}^M > 0$, it is clear that this condition is satisfied if and only if

$$0 < c_j \tag{82}$$

$$0 < \psi_j < 1 \tag{83}$$

$$0 < \kappa_0^{M,j} \tag{84}$$

Equations 82 follows from the fact that $(1-\psi_j)\kappa_0^{M,j} - \psi_j \bar{e}^M) > 0$ because $0 < \psi_j < 1$ and $e^M < \kappa_0^{M,j}$. Equation 82 - 83 are always satisfied given the model's parameter assumptions. 83 is satisfied because $\eta > 0$. The lower dominance region in which the MF run is the unique equilibrium outcome always exists and is non-empty.

Upper-dominance region $[\overline{x_j}, \infty)$. Let $\bar{e}^M = E[e^M]$ be the expected mass of impatient (early) MF investors at $t = 1$ as before and $\bar{l}^M = E[l^M]$ the expected mass of patient (late) MF investors. $k = \kappa_0^{M,j} - \eta \leq 1$ denotes the total mass of non-sleepy MF investors and is known after initial allocation decisions have been made at $t = 0$. Using equation 20 and lemma 2, the upper dominance region with respect to the fundamental x_j is defined by the following condition:

$$E[P_2^j - \Delta_2^{M,j} - P_1^{M,j} R^f | x_j = \overline{x}_j \cup \Delta\kappa_1^{M,j} = k] = 0 \tag{85}$$

Equation 85 is similar to the condition for the lower dominance region, $\underline{x}_j$, in equation 72. The distinguishing feature between both regions is the conditioning information in the expectations operator.

Using $P_2^j = x_j$ and $x_j > 0$, from 85 follows

$$0 = x_j - x_j E\left[\frac{\left(\frac{\Delta\kappa_1^{M,j}(P_1^{M,j} - P_{1^+}^j)}{P_{1^+}^j} \right)}{\kappa_0^{M,j} - \Delta\kappa_1^{M,j}} \middle| x_j = \overline{x}_j \cup \Delta\kappa_1^{M,j} = \kappa^M \right] - \psi_j x_j R^f. \tag{86}$$

Substituting $P_{1^+}^j$ from the index market clearing at $t = 1^+$ following proposition 3:

$$0 = 1 - \frac{\left(\frac{\Delta\kappa_1^{M,j}(\psi_j \overline{x}_j - \frac{1}{2}(\overline{x}_j + \sqrt{\overline{x}_j^2 - 4c_j \ \psi_j \ \overline{x}_j \ \Delta\kappa_1^{M,j}})) }{\frac{1}{2}(\overline{x}_j + \sqrt{\overline{x}_j^2 - 4c_j \ \psi_j \ \overline{x}_j \ \Delta\kappa_1^{M,j}})} \right)}{\kappa_0^{M,j} - k} - \psi_j R^f. \tag{87}$$

Using the fact that $\kappa_0^{M,j} = k + \eta$ and $R^f = 1$,

$$1 - \psi_j = \frac{\left(\frac{k(\psi_j \overline{x}_j - \frac{1}{2}(\overline{x}_j + \sqrt{\overline{x}_j^2 - 4c_j \ \psi_j \ \overline{x}_j \ k})) }{\frac{1}{2}(\overline{x}_j + \sqrt{\overline{x}_j^2 - 4c_j \ \psi_j \ \overline{x}_j \ k})} \right)}{\eta}. \tag{88}$$

Simplifying:

$$\eta(1-\psi_j) = \frac{2k\psi_j}{1+\sqrt{1-\frac{4c_j\ \psi_j\ k}{\overline{x}_j}}} - k. \tag{89}$$

Solving for $\overline{x}_j$:

$$\overline{x}_j = \frac{4c_j\psi_j k\ \left(\eta(1-\psi_j)+k\right)^2}{\left(\eta(1-\psi_j)+k\right)^2 - \left((2\psi_j-1)k - \eta(1-\psi_j)\right)^2}. \tag{90}$$

Finally, substituting $k = \kappa_0^{M,j} - \eta$ the upper dominance region is defined by:

$$\overline{x}_j = \frac{c_j\left(\kappa_0^{M,j} - \psi_j\eta\right)^2}{(1-\psi_j)\kappa_0^{M,j}}. \tag{91}$$

Existence of upper dominance region. For the upper dominance region to be non-empty it must hold that $\overline{x}_j > 0$ (necessary condition). Besides, $\underline{x}_j \leq \overline{x}_j$ (sufficient condition). Given the lower dominance region exists, that is condition 81 is satisfied, if $\underline{x}_j \leq \overline{x}_j$ holds, the upper dominance region must exist.

Necessary condition: The numerator and denominator of equation 91 must both be strictly positive. First, the numerator of equation 91,

$$c_j\left(\kappa_0^{M,j} - \psi_j\eta\right)^2 > 0, \tag{92}$$

is always strictly positive because $\kappa_0^{M,j} > \eta$ and $0 < \psi < 1$. Second, for the same reasons the denominator of equation 91 is also always strictly positive:

$$(1-\psi_j)\kappa_0^{M,j} > 0. \tag{93}$$

Hence, $\overline{x}_j > 0$.

Sufficient condition: $\underline{x}_j \leq \overline{x}_j$ requires:

$$c_j\left(\psi_j\bar{e}^M + (1-\psi_j)\kappa_0^{M,j}\right)^2 \leq c_j\left(\kappa_0^{M,j} - \psi_j\eta\right)^2. \tag{94}$$

For $c_j > 0$, this simplifies into:

$$\bar{e}^M + \eta \leq \kappa_0^{M,j},$$

which is always satisfied by definition because $\kappa_0^{M,j} = e^M + l^M + \eta$.

Proof of Corollary 6. From equation 80 the lower dominance region is defined by:

$$\underline{x}_j = \frac{c_j(\psi_j \bar{e}^M + (1-\psi_j)\kappa_0^{M,j})^2}{(1-\psi_j)\kappa_0^{M,j}}, \tag{95}$$

where $\kappa_0^{M,j} = \int_i \theta_0^{i,M,j} di + \eta$. Let $k = \int_i \theta_0^{i,M,j} di$. Taking partial derivatives with respect to the key model parameters and variables gives:

$$\frac{\partial \underline{x}_j}{\partial c_j} = \frac{(\psi_j \bar{e}^M + (1-\psi_j)\kappa_0^{M,j})^2}{(1-\psi_j)\kappa_0^{M,j}} > 0, \tag{96}$$

$$\frac{\partial \underline{x}_j}{\partial \bar{e}^M} = \frac{2c_j\psi_j(\psi_j \bar{e}^M + (1-\psi_j)\kappa_0^{M,j})}{(1-\psi_j)\kappa_0^{M,j}} > 0 \tag{97}$$

$$\frac{\partial \underline{x}_j}{\partial \eta} = c_j\left((1-\psi_j) - \frac{(\psi_j \bar{e}^M)^2}{(1-\psi_j)(k+\eta)^2}\right). \tag{98}$$

Where equation 98 simplifies into:

$$\begin{aligned} & c_j\left((1-\psi_j) - \frac{(\psi_j \bar{e}^M)^2}{(1-\psi_j)(k+\eta)^2}\right) \\ & = \underbrace{\frac{c_j}{(1-\psi_j)(k+\eta)^2}}_{>0}\left(\underbrace{(1-\psi_j)^2(k+\eta)^2}_{>0} - \underbrace{(\psi_j \bar{e}^M)^2}_{>0}\right) \end{aligned} \tag{99}$$

The sign of the partial derivative $\frac{\partial \underline{x}_j}{\partial \eta}$ depends on the relative magnitude of $(1-\psi_j)^2(k+\eta)^2$ and $(\psi_j \bar{e}^M)^2$. Specifically, for equation 98 it follows that $\frac{\partial \underline{x}_j}{\partial \eta} < 0$ whenever $(1-\psi_j)(k+\eta) < \psi_j \bar{e}^M$. By assumption $0 < \psi_j < 1$ is close to 1, while $9 \leq k \leq 1$. Therefore, as long as η is relatively small and the mass of early MF investors is large relative to the late (patient) MF investors,

$$\eta < \underbrace{\frac{\psi_j}{(1-\psi_j)}}_{>1} \bar{e}^M - \underbrace{k}_{\geq \bar{e}^M}, \tag{100}$$

the size of the lower dominance region decreases as the mass of sleepy MF investors increases, $\frac{\partial \underline{x}_j}{\partial \eta} < 0$.

Proof of Corollary 7. From equation 91 the upper dominance region is defined by:

$$\overline{x}_j = \frac{c_j\left(\kappa_0^{M,j} - \psi_j \eta\right)^2}{(1-\psi_j)\kappa_0^{M,j}}. \tag{101}$$

where $\kappa_0^{M,j} = \int_i \theta_0^{i,M,j} di + \eta$. Let $k = \int_i \theta_0^{i,M,j} di$. When $\overline{x}_j$ increases, the size of the no-run region decreases because $\overline{x}_j$ marks the lower bound of this region.

Taking partial derivatives with respect to the key model parameters and variables gives:

$$\frac{\partial \underline{x}_j}{\partial c_j} = \frac{\left(\kappa_0^{M,j} - \psi_j \eta\right)^2}{(1-\psi_j)\kappa_0^{M,j}} > 0, \tag{102}$$

$$\frac{\partial \underline{x}_j}{\partial \kappa_0^{M,j}} = \frac{c_j((\kappa_0^{M,j})^2 - \psi_j^2 \eta^2)}{(1-\psi_j)(\kappa_0^{M,j})^2} > 0, \tag{103}$$

$$\frac{\partial \underline{x}_j}{\partial \eta} = \frac{c_j}{1-\psi_j}\left(1 + 2\psi_j\left(\frac{\psi_j}{(k+\eta)^2} - 1\right)\right) \tag{104}$$

Equations 102 and 103 follow from the facts that $\kappa_0^{M,j} = k + \eta$ and $0 < \psi_j < 1$. Equation 104 implies that the size of the no-run region increases with the mass of sleepy investors, $\frac{\partial \underline{x}_j}{\partial \eta} < 0$, when the following condition is satisfied:

$$0.5 < \psi_j - \frac{\psi_j^2}{(k+\eta)^2}. \tag{105}$$

Proof of Corollary 8. This is a direct result from equations 29 and 30 for $c_j = 0$.

Proof of Lemma 5. The expected ETF and MF payoffs of a certain long-term investor, characterized by $\lambda_i = 0$ as of $t = 0$, are given by:

$$\begin{aligned}
E_0[w_1^i | \theta_0^{i,E,j} = 1] &= x_j \\
E_0[w_1^i | \theta_0^{i,M,j} = 1] &= \psi\left(\int_0^{\underline{x}_j} x_j dx + \frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j} x_j dx\right) \\
&\quad + \left(\frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j} (x_j - \Delta_2^{M,j}) dx + \int_{\overline{x}_j}^{\infty} (x_j - \Delta_2^{M,j}) dx\right).
\end{aligned}$$

Since $0 < \psi < 1$ and $\Delta_2^{M,j} \geq 0$, $E_0[w_1^i | \theta_0^{i,E,j} = 1] > E_0[w_1^i | \theta_0^{i,M,j} = 1]$, so investors with $\lambda_i = 0$ always choose the ETF.

The expected ETF and MF payoffs of a certain short-term investor, characterized by $\lambda_i = 1$ as of $t = 0$, are given by:

$$\begin{aligned}
E_0[w_1^i | \theta_0^{i,E,j} = 1] &= \mu_j - \bar{e}^M(c_j + \phi_j), \\
E_0[w_1^i | \theta_0^{i,M,j} = 1] &= \psi \mu_j.
\end{aligned}$$

Under the given parameter assumptions $(1-\psi)\mu_j < \bar{e}^M(c_j+\phi_j)$, it holds that investors with $\lambda_i = 1$ always choose the MF because $E_0[w_1^i|\theta_0^{i,M,j}=1] > E_0[w_1^i|\theta_0^{i,E,j}] = 1$.

It is noteworthy that in the equilibrium with imperfectly liquid index markets the mass of ETF investors, and therefore the ETF's AUM, must be strictly positive, $\kappa_0^{E,j} > 0$. Otherwise, a contradiction arises: If all investors with $\lambda_i > 0$ invested in the MF, the mass of ETF investors would be infinitely small as only investors with $\lambda_i = 0$ would invest in ETFs, so in the limit $\kappa_0^{M,j} \to 1$ and $\kappa_0^{E,j} \to 0$. In this limiting case the ETF would perfectly track the index at all times: $P_1^{E,j} \to x_j$ and $P_2^{E,j} = x_j$, and therefore dominate the MF because $\psi < 1$ and $\Delta_2^{M,j} \geq 0$. As a result, investors at the margin, characterized by $\lambda_i = \epsilon$ where $\epsilon \to 0$, would optimally choose to deviate and switch to the ETF. Hence, $\kappa_0^{E,j} = 0$ cannot be an equilibrium. It follows that the initial mass of ETF investors must be non-zero, $\kappa_0^{E,j} > 0$. Given $\lambda_i \sim U[0,1]$, the law of large numbers then implies that the expected mass of impatient ETF investors must also be non-zero $\bar{e}^E > 0$.

Proof of Proposition 5. A challenge emerges because an investor's expected ETF and MF payoffs as of $t=0$ depend on the mass and characteristics of other fund investors. I conjecture that investors at $t=0$ invest according to the strategy $\{\theta_0^{i,E,j}, \theta_0^{i,M,j}\} = \{1,0\}\ \forall i$ with $\lambda_i \leq \lambda'$ and $\{\theta_0^{i,E,j}, \theta_0^{i,M,j}\} = \{0,1\}\ \forall i$ with $\lambda_i > \lambda'$ for some $\lambda' \in (0,1)$ and show that any individual investor with $\lambda_i \in [0,1]$ has no incentive to deviate from this strategy profile.

Following lemma 5, it must hold that $0 < \kappa_0^{E,j} < 1$ and $0 < \kappa_0^{M,j} < 1$. The cross-sectional distribution of liquidity risks $\lambda_i \sim U[0,1]$ together with the law of large numbers then imply that the expected mass of impatient ETF and MF investors is non-zero in equilibrium, $0 < \bar{e}^E < 1$ and $0 < \bar{e}^M < 1$. Therefore the cut-off liquidity risk level must satisfy:

$$0 < \lambda' < 1$$

Next, for the conjectured equilibrium to exist, holding all else equal, the expected ETF payoff must be decreasing in λ_i while the expected MF payoff must be increasing in λ_i. First, the expected ETF payoff as a function of λ_i, here denoted by $f(\lambda_i) = E_0[\lambda_i P_1^{E,j} + (1-\lambda_i)P_2^{E,j}|\lambda_i, \lambda']$, follows from equation 33:

$$f(\lambda_i) = \mu_j - \lambda_i \bar{e}^E(c_j+\phi_j),$$

where $\bar{e}^E = \frac{\lambda'}{2} > 0$.

Partial differentiation with respect to λ_i gives:

$$\frac{\partial f(\lambda_i)}{\partial \lambda_i} = -\bar{e}^E(c_j+\phi_j) < 0.$$

For $c_j + \phi_j > 0$ (baseline assumption), the expected ETF payoff strictly decreases in λ_i.

Second, the expected MF payoff as a function of λ_i, here denoted by $g(\lambda_i) = E_0[\lambda_i P_1^{M,j} + (1-\lambda_i)P_2^{M,j}|\lambda_i, \lambda']$, follows from equation 34:

$$g(\lambda_i) = \lambda_i \psi \mu_j + (1-\lambda_i)\psi\Big(\int_0^{\underline{x}_j} x_j dx + \frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j} x_j dx\Big) + (1-\lambda_i)\Big(\frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j}(x_j - \Delta_2^{M,j})dx + \int_{\overline{x}_j}^{\infty}(x_j - \Delta_2^{M,j})dx\Big).$$

Note that $\Delta\kappa_1^{M,j} = \bar{e}^M = \frac{1+\lambda'}{2}$.

Using the fact that $\lambda_i \perp\!\!\!\perp x_j$, and the expression for the tracking difference from equation 17 together with the equilibrium index price at $t = 1$ from proposition 3, partial differentiation with respect to λ_i gives:

$$\frac{\partial g(\lambda_i)}{\partial \lambda_i} = \psi\mu_j - \psi\Big(\int_0^{\underline{x}_j} x_j dx + \frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j} x_j dx\Big) - \Big(\frac{1}{2}\int_{\underline{x}_j}^{\overline{x}_j}(x_j - \Delta_2^{M,j})dx + \int_{\overline{x}_j}^{\infty}(x_j - \Delta_2^{M,j})dx\Big) \tag{106}$$

Using the expression for the terminal MF tracking difference as a function of x_j from equation 45 (lemma 2) and noting that $\Delta\kappa_1^{M,j} = \bar{e}^M$ over the no-run region of x_j, implies $\frac{\partial g(\lambda_i)}{\partial \lambda_i} > 0$.

Proof of Proposition 6. By definition the optimal swing factor is such that it fully eliminates the externalities between redeeming and remaining MF investors, and therefore any first-mover advantage. Formally $s_j{}^*$ must ensure that the price at which early MF investors redeem is equal to the price at which the MF itself trades in index markets at $t = 1^+$:

$$P_1^{M,j} = P_{1^+}^j.$$

Given MF outflows of $\kappa_1^{M,j}$, the index price at $t = 1^+$ is:

$$P_{1^+}^j = x_j - c_j\kappa_1^{M,j}.$$

Hence, the swing factor $s_j = c_j\Delta\kappa_1^{M,j}$ is optimal because $x_j - s_j{}^* = P_{1^+}^j$.

Proof of Corollary 9. With optimal swing pricing the MF NAV at $t = 1$ is equal to the index price the MF trades at once it passes on fund flows to index markets, $P_1^{M,j,Swing} = P_{1^+}^j$. This follows from proposition 6. Therefore, the MF perfectly tracks its index at $t = 1$. At the same time, corollary 1 implies that the MF tracking difference at $t = 2$ is zero, $\Delta_2^{M,j} = 0$, if the MF is perfectly forward looking at $t = 1$. As a result $P_1^{M,j,Swing} = P_{1^+}^j$ implies

$P_2^{M,j,Swing} = P_2^j$ irrespective of MF flows at $t = 1$. The MF perfectly replicates the index at all times under the optimal swing pricing rule.

While the payoff of the ETF and the MF with swing pricing is the same at $t = 2$, the ETF remains subject to mispricing risk over the short term, $P_1^{E,j} = x_j - \Delta\kappa_1^E(c_j + \phi_j)$. Therefore, no short- or long-term MF investor has any incentive to deviate and invest in the ETF.

Proof of Corollary 10. This follows directly from proposition 6 and corollary 9 for $\Delta\kappa_1^{M,j} > 0$.

D List of model variables

Table D.1: Model notation

Exogenous parameters and their interpretation.	
Parameter	Definition
x_j	State variable / terminal index payoff
λ_i	i's probability of a liquidity shock at $t = 1$
ϕ_j	AP balance sheet capacity constraint
c_j	Index market maker inventory cost
η	Mass of sleepy investors
R^f	Risk-free rate, normalized to equal one
Endogenous quantities and their interpretation.	
Parameter	Definition
κ_t^E	Share (mass) of ETF investors at t
κ_t^M	Share (mass) of MF investors at t
$\Delta\kappa_1^E$	Volume of ETF liquidations (outflows) at $t = 1$
$\Delta\kappa_1^M$	Volume of MF redemptions at $t = 1$
P_t^j	Index market price at t
P_t^E	ETF market price
NAV_t^E	ETF net asset value
NAV_t^M	Mutual fund net asset value, $P_t^M = NAV_t^M$
ϵ_t^E	Relative ETF mispricing (discount), $\epsilon_t^E = P_t^j - P_t^E$
$\theta_t^{i,M}$	Units of MF shares held by agent i at t
$\theta_t^{i,E}$	Units of ETF shares held by agent i at t