Exchange-Traded Funds: An Overview of Institutions, Trading, and Impacts

Ananth Madhavan

Global Head of iShares® Research at BlackRock® Inc., San Francisco, California 94105; email: Ananth.Madhavan@blackrock.com

Annu. Rev. Financ. Econ. 2014. 6:311–41

The *Annual Review of Financial Economics* is online at financial.annualreviews.org

This article's doi:
10.1146/annurev-financial-110613-034316

Keywords

ETF, investment management, price dynamics, systemic risk

Abstract

Exchange-traded funds (ETFs) have grown substantially in diversity and size in recent years, reflecting a broader shift toward passive, index investing. As a consequence, there is increased attention by investors, regulators, and academics seeking to assess and understand the implications of this rapid growth. This article provides a unified framework to examine these issues and review the research to date, demonstrating that ETFs have extended significant benefits to investors and to the functioning of markets that meaningfully outweigh any perceived or actual weaknesses.

1. INTRODUCTION

Exchange-traded funds (ETFs) have grown substantially in size, diversity, and market significance in recent years, reflecting a broader interest in passive, index investing. Assets in US ETFs rose from $70 billion in 2000 to $1.7 trillion by mid-2014, with worldwide totals in all asset classes now exceeding $2.5 trillion.[1] As a consequence, there is increased attention by investors, regulators, and academics seeking to assess and understand the implications of this rapid growth. This article provides an integrated framework to address questions about ETFs and, more broadly, passive indexing.

The common characteristic of ETFs is that they are traded intraday on an organized exchange. Although simple, this description ignores many important differences between ETFs and primary securities such as stocks or bonds that have important implications for investors. I develop a model that emphasizes arbitrage as a driver of liquidity and price dynamics. I use the framework to analyze questions concerning price discovery, the dynamics of premiums and discounts, return autocorrelations, performance and tracking relative to benchmark, transaction costs, and liquidity sourcing in underlying and secondary markets. I also use this framework to examine questions about "alternative beta" based on quantitative portfolio construction techniques and active funds that attempt to outperform a benchmark index. Finally, I review several recent policy questions concerning the impact of passive flows on underlying securities, leveraged and inverse ETFs, the role of ETFs in the Flash Crash of May 2010, and issues concerning systemic risk including excess shorting, settlement failures, etc. I conclude that ETFs have extended significant benefits to investors and to the functioning of markets and that concerns over their operation are largely due to misconceptions about pricing and liquidity (for a comprehensive overview of the benefits of ETF ownership that dispels many of the myths concerning index products more generally, see Golub et al. 2013).

This article is organized as follows: Sections 2 and 3 provide the required institutional background necessary to understand the operation of ETFs. Section 4 develops a canonical model used to analyze ETF premiums and discounts, price discovery, volatility, liquidity, and transaction costs. Sections 5 and 6 extend this model to examine issues related to active ETFs. Sections 7 and 8 consider current issues including concerns about the impact of ETF flows on underlying markets, about systemic risk and the Flash Crash, and about the impact of leveraged and inverse ETFs. Finally, Section 9 concludes with a focus on the role of public policy and regulation and a review of the value of ETFs as an investment vehicle.

2. ACTIVE AND PASSIVE INVESTING

2.1. Growth of Indexing

Passive investing, where investors seek to match broad market capitalization–weighted indexes, has grown tremendously in the past few decades. Sullivan & Xiong (2012) note that, although passively managed funds represent only approximately one-third of all fund assets, their average annual growth rate since the early 1990s is 26%, double that of actively managed assets. In part, this growth reflects the establishment of modern financial economics including the value of broad portfolio diversification and the concept that prices are informationally efficient. Other important considerations in the growth of indexing include growing awareness of transaction costs, taxes, and lower fees and turnover relative to active management.

[1]ETFs are a subset of a broader group of investment vehicles termed exchange-traded products (ETPs). In an ETF, the underlying basket securities are physically represented with the objective of mimicking the performance of a broad market index. Exchange-traded notes, by contrast, are senior, unsecured, and uncollateralized debt exposed to credit risk. Some ETPs contain embedded leverage and are, thus, quite different from conventional, physically based ETFs.

By contrast, active investing, as defined here, requires both an investor to depart from market capitalization–based weighting and a conscious decision on what securities or asset classes to over- or underweight relative to the broader market. Active management is thus a zero-sum game absent fees. Furthermore, evidence indicates it is difficult to identify skilled active managers. Index investing, through the wrappers of index mutual funds and ETFs, offers low-cost, diversified exposure to various market segments.

2.2. Investment Strategies

An important theme throughout this article, and an organizing principle, is that the range of investing strategies from passive to active is a continuum of exposures and different degrees of transparency into the underlying holdings (see **Table 1**). I distinguish between passive/exposure-based funds (against a benchmark or without an index), transparent active funds (where the weighting scheme is not market capitalization based but portfolio construction is model based), and nontransparent active funds (including fundamental funds).

Most index-tracking ETFs and open-ended funds mimic standard indexes that weight components based on their market capitalization. This weighting scheme is sensible, offering lower turnover and greater liquidity (or investability) relative to many alternatives. Even pure passive indexing against a common benchmark such as the S&P 500 requires active decisions on the part of the investor and index provider. For example, index membership is selected by a committee chosen by Standard & Poor on the basis of a variety of criteria, both objective and subjective, and the investor needs to consider whether to allocate funds across other capitalization segments or asset classes and, if so, in what proportions. Some funds, however, track benchmarks that are expressions of "active" investment strategies based on systematic, transparent model-driven rules. So-called alternative beta includes several important subcategories, as discussed below in Section 5 (see also Chow et al. 2011).

3. INVESTMENT VEHICLES

3.1. Exchange-Traded Funds and Mutual Funds

So far, investment strategies have been discussed in terms of the continuum from passive to active without explicit differentiation among the types of funds (or wrappers) used to achieve these objectives. A point worth emphasizing is that ETFs are more complex than commonly believed and possess several unique features that distinguish them from other fund structures. They are not simply exchange-traded versions of index mutual funds. Rather, ETFs share elements of both open- and closed-end mutual funds. Relative to index mutual funds, ETFs add the ability to trade intraday and the tax advantages that derive from the ability to use in-kind transfers, as explained below. ETFs are similar to open-ended funds in that shares can be created or redeemed at the end of the trading day for the current per-share net asset value (NAV) of the fund. Unlike open-ended funds, however, ETFs issue and redeem shares at NAV only in a minimum size (creation unit) and only with market-making firms known as authorized participants (APs). Also, unlike open-ended funds, but similar to closed-end funds, ETFs are tradable intraday on an exchange in the secondary market at prices that can and often do vary from NAV.

3.2. Diversity of Exposures

Exchange-traded products (ETPs) provide exposure to a wide range of asset classes (e.g., equities, fixed income commodities, and currencies), strategies (e.g., passive index, model based), and regions. **Figure 1** shows the breakdown of global ETP assets under management by asset class,

Table 1 Continuum of investment strategies from most passive to most active

Most passive → Most active

Exposure based		Transparent active		Nontransparent active
Pure index	**Exposure without an index**	**Alternatively weighted index**	**Model driven**	**Fundamental**
Reflects the performance of an asset class using a market cap index	Passively reflects the performance of an asset class without using an index process	Delivers on a specific enhancement strategy expressed as an alternative weighted index	Seeking to generate alpha or controlled risk exposures using quantitative models	Seeking to generate alpha without explicit portfolio formation rules
The majority of stock and bond index funds and exchange-traded funds, either exactly replicating the index or using optimization	Optimized for tax efficiency or other goals	Fundamental indexes based on earnings, etc., that are not market capitalization weighted	Minimum-volatility funds, factor funds, and risk-weighted funds	Either bottom-up security selection or liquidity selection

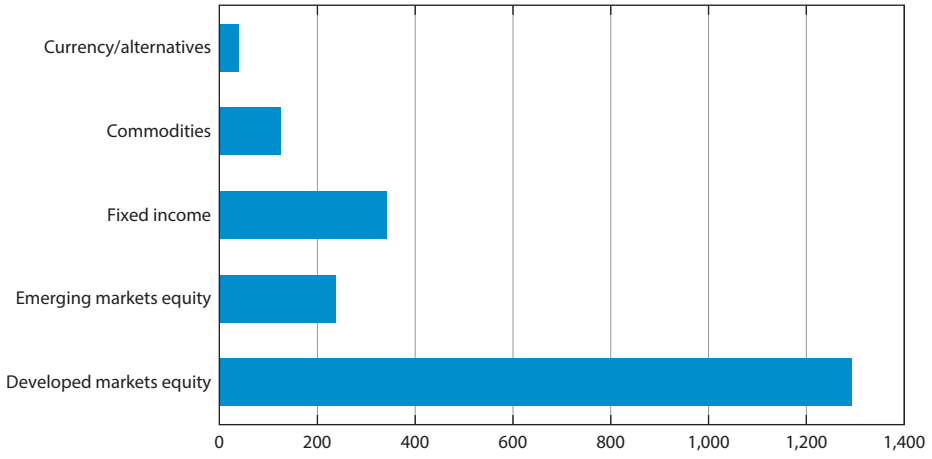

Figure 1

Global assets under management in US dollars (billions). Source: BlackRock ETP Landscape; based on data as of June 2013.

amounting to $2,038 billion, as of June 2013. By way of comparison, in 2005, global assets under management were $428 billion with 87% in the developed markets equity category as opposed to 63% in 2013. Over this period, the growth of nontraditional asset classes including emerging markets equity, fixed income, commodities, and currencies (in **Figure 1**, this category also includes alternatives and asset allocation) has been especially dramatic.

There are also important differences in fund structure. Index-tracking ETFs typically hold a physical portfolio of securities that closely resembles, but does not necessarily fully replicate, their benchmark index. So-called synthetic funds, however, track an index by holding derivatives or swaps to replicate the performance of their benchmarks. Here, the term ETF refers to open-end funds that are exchange traded and invest in a portfolio of physical securities.

3.3. Creation-Redemption Mechanism

A unique feature of ETFs is the creation-redemption mechanism, which relies on arbitrage. Transactions between an ETF manager and an AP are typically either for cash or "in-kind" where the AP delivers or receives a basket of securities identical (or very similar) to the ETF's holdings.[2] Current fund holdings and the basket of securities the ETF is willing to accept for in-kind creations or redemptions the next business day are published at the end of each trading day. That transparency is important when we consider active ETFs, an increasingly popular innovation. As with all investors, APs can buy or sell ETF shares in the secondary market exchange, but they also can purchase or redeem shares directly from the ETF if they believe there is a profit opportunity. ETF shares are created or redeemed at NAV at the end of each trading day. However, it is important to understand that the AP can, and typically will, lock in its profits during the day by selling the higher-priced asset while simultaneously buying the lower-priced asset. The creation-redemption mechanism works through arbitrage to keep the ETF price within a transaction cost bound to the

[2]Unlike open-ended funds, in the case of cash redemptions, transaction charges resulting from investing or raising the cash are absorbed by the AP and not the ETF. Cash redemptions may be required because some ETF holdings, such as certain emerging market stocks, are subject to legal restrictions that prevent in-kind transfers.

fair value of an ETF's holdings in the underlying market. The arbitrage mechanism also encourages APs to provide offsetting liquidity when there is an excess of buying or selling demand for ETF shares. In addition to the above, secondary market liquidity is another distinguishing feature of ETFs relative to mutual funds that offer liquidity only at the end of the day. However, the traded volume of an ETF can be a misleading metric of its true liquidity, as the traded volume overlooks the liquidity of the underlying constituents of the ETF (as discussed below).

Example 1: iShares Emerging Markets Volumes and Creation Activity For many ETFs, investors can buy or sell the fund in the secondary market without any creations or redemptions. Consider the iShares Emerging Markets ETF (EEM), whose constituents are often illiquid and difficult to access but which is among the most liquid US-listed equities with average daily volume of ~56 million shares in mid-August 2013. Yet, from August 6 to August 13, 2013, ~360 million shares of EEM were traded with no net redemptions or creations.

4. CONCEPTUAL FRAMEWORK

4.1. Definitions

This section develops a canonical model that is used throughout this article and that builds on the notion of arbitrage. Time is measured in fixed calendar intervals Δt (such as seconds, minutes, or days) and the time index is denoted by $t = 0,1,2$, etc. The individual asset in the ETF basket is denoted by $i = 1, \ldots, N$. **Table 2** lists the variables and parameters used throughout.

Table 2 Variables and parameter definitions

Variable	Definition
p_t	Execution price of the ETF at the end of period t
n_t	NAV of the ETF in period t
m_t	Midquote price of the ETF in period t
v_t	Expected value of the ETF in period t conditional on public information
r_t	Innovation in expected value from period $t-1$ to t
π_t	ETF premium in period t
q_t	Signed volume in ETF in period t
S_t	Index value in period t
o_t	Shares outstanding of ETF in period t
u_t	Deviation of price from expected value in period t
h	Security holding weights of the ETF, an $N \times 1$ vector
h_b	Benchmark security holding weights of the underlying index, an $N \times 1$ vector
α	Forecast alpha (excess return), an $N \times 1$ vector
σ	Volatility or standard deviation of returns
V	Covariance matrix, an $N \times N$ matrix
φ	Staleness parameter ($0 \leq \varphi \leq 1$)

We can interpret the price variables as inclusive of corporate actions and dividends. Furthermore, if they are represented in natural logs so that first differences are continuously compounded returns, then the change (or innovation) in expected values conditional upon public information over the interval from $t-1$ to t is denoted by $r_t = v_t - v_{t-1}$. The return over longer periods is the sum of interval returns: From period 0 to T, the total return is $r_{0,T} = \sum_{t=1}^{T} r_t = v_T - v_0$.

4.2. Pricing and Liquidity

We begin by modeling the evolution of prices, conditional values, and premiums, all topics that are fundamentally linked. We assume the ETF price is the (unobserved) expected value of the asset conditional upon public information, v_t, plus a noise shock (also unobserved) u_t that captures microstructure noise and transitory liquidity effects:

$$p_t = v_t + u_t. \tag{1}$$

For a domestic, actively traded equity index, a reasonable proxy for expected value v_t is the weighted midquote (average of the bid and offer prices) of the basket securities. However, for international or less liquid baskets (e.g., fixed income or small cap equity), there may not be good observable proxies for expected value, so Equation 1 cannot be used directly to assess the impact of noise or liquidity effects. Conditional expectations follow a martingale, meaning their changes $r_t = v_t - v_{t-1}$ over the interval from $t-1$ to t are a surprise or innovation. We implicitly assume investors have homogeneous information, although some (noise) traders may trade for noninformational reasons.

To examine premiums and discounts, we need to define the NAV formally for the fund. The NAV is determined by a market vendor using last-recorded prices, perhaps adjusted using a proprietary process. I denote the NAV by n_t, defined as the weighted average of the last recorded prices of the basket securities. The last recorded price (or quote) may be stale, i.e., because asset i does not trade or no quotes are available. In this case, the last recorded quote is, on average, the previous period's expected value, and so on back in time. Accordingly, the fund's NAV is modeled as the weighted average of past expected values:

$$n_t = (1-\varphi)\sum_{j=0}^{\infty} \varphi^j v_{t-j}. \tag{2}$$

In this formulation, the weight placed on the price or quote $j-$ lags ago is $(1-\varphi)\varphi^j$, where $0 \le \varphi \le 1$. The weights on past values sum to 1 and decay exponentially. Here $\varphi(\Delta t)$ captures staleness and depends implicitly on the calendar interval Δt chosen. For example, if time is measured in, say, intervals of 15 s, staleness may be larger than if measured in hourly or daily intervals. It is straightforward to show from Equation 2 that NAV is a weighted average of the current expected value and past NAV:

$$n_t = (1-\varphi)v_t + \varphi n_{t-1}. \tag{3}$$

Equation 3 is intuitively appealing. For example, an international fund that is traded in the United States may have up-to-date quotes for only a fraction of its constituents (e.g., cross-listed firms) so the NAV is a weighted average of current quotes and past NAV. An error term could be added to the formula for the NAV to capture pricing errors, but this does not materially change the analysis.

The premium (or discount) is defined as the deviation of the ETF price from the NAV:

$$\pi_t = p_t - n_t. \tag{4}$$

If the basket weights are equal to the benchmark weights, so that $h = h_b$, the return on the ETF is equal to the return on the NAV of the benchmark plus the change in the premium. By contrast, for an open-ended fund, transactions occur only at the end of the day and at the NAV. The fact that the NAV can be stale provides opportunities to trade on stale prices, possibly diluting value for existing shareholders. This so-called market-timing issue is especially important for international funds and has led those funds to adopt fair value pricing, where the NAV is adjusted on the basis of a variety of factors. For example, Grégoire (2013) notes that despite direction by the Securities and Exchange Commission in 2004, there is still evidence that mutual funds do not fully adjust their valuations and returns remain predictable.

4.2.1. Arbitrage. It is important to understand that the profit of a market maker is not the premium but the deviation of price from expected value, u_t. Whereas creations and redemptions are at the NAV, this is really an end-of-day accounting or book-entry transaction. When price is above expected value (plus a premium for transaction costs, taxes, commissions, and fees, etc.), the market maker can sell the ETF while simultaneously purchasing the basket of securities (or obtaining that exposure through some other mechanism such as a swap) to make an expected profit. Of course, the market maker or other investor may choose not to hedge their position but simply sell the ETF and carry some inventory risk from an unhedged position. Assuming the capital of market makers is finite (or that they are risk averse or face market impact costs), price does not instantly revert back to expected value (net transaction costs), but rather corrects over time as in microstructure models of market impact. So, given the buying or selling flow by arbitrageurs (APs or other investors) and noise traders, the deviation from expected value is modeled as

$$u_t = \psi u_{t-1} + \varepsilon_t, \tag{5}$$

where ε_t is a mean zero error term with variance τ^2 that captures the effect of noise traders (and other random factors) and ψ is an inverse measure of the speed of error correction through arbitrage. Intuitively, arbitrageurs act to correct pricing errors, and their trading causes a convergence of price to expected value.

4.2.2. Return volatility. There is often considerable confusion regarding the returns of ETFs versus the index. We can use the framework above to clarify our thinking around performance and volatility. Returns on the NAV are

$$n_t - n_{t-1} = (1 - \varphi) r_t + \varphi(n_{t-1} - n_{t-2}). \tag{6}$$

Assuming the change in expected value is independently and identically distributed, $\sigma^2(r_t) = \sigma^2$, we obtain the variance of NAV returns as

$$\sigma^2(n_t - n_{t-1}) = \sigma^2 \frac{(1 - \varphi)^2}{1 - \varphi^2}. \tag{7}$$

So, NAV return variance is a fraction of fundamental variance $\sigma^2(1 - \varphi)/(1 + \varphi)$, which is strictly less than the variance of fundamental return if the staleness parameter φ is in the interval (0,1).

Turning now to the variance of ETF price returns, the price-based return on the ETF is $(p_t - p_{t-1}) = (v_t - v_{t-1}) + (u_t - u_{t-1})$. As the first term is the return innovation, the variance of ETF price returns can be shown to be

$$\sigma^2(p_t - p_{t-1}) = \sigma^2 + 2\tau^2(1+\psi)^{-1}. \tag{8}$$

In this case, let L denote the lag operator, i.e., $Lu_t = u_{t-1}$. Then, $u_t = (1-\psi L)^{-1}\varepsilon_t = \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \ldots$. The shock change $(u_t - u_{t-1}) = \varepsilon_t + (\psi - 1)(1-\psi L)^{-1}L\varepsilon_t$, so $\sigma^2(u_t - u_{t-1}) = \tau^2\left(1 + \frac{(1-\psi)^2}{1-\psi^2}\right)$, which can be simplified to $2\tau^2(1+\psi)^{-1}$. The first term in Equation 8 is fundamental return variance and the second reflects variance from bid-ask bounce and liquidity shocks.

So, we obtain a key result: ETF return volatility will exceed that of NAV returns. The differences are greater the higher the degree of staleness and the larger the variance of liquidity and other random shocks. Over longer intervals, the return variance will scale with time, so for an interval T we have $\sigma^2(r_{0,T}) = \sigma^2 T$. The variance of the microstructure shock difference component of ETF returns, $\sigma^2(u_t - u_{t-1})$, is stationary and does not scale with time. As time increases, the effect of staleness (recall the parameter φ depends on the calendar units of measurement) will diminish and NAV and ETF return differences narrow over longer intervals.

4.2.3. Autocorrelation. The autocorrelation of ETF and NAV returns is also quite different. Using the definition of NAV, we have $n_t - n_{t-1} = (1-\varphi)r_t + \varphi(n_{t-1} - n_{t-2})$, so the covariance of successive NAV returns is

$$E\left[(n_t - n_{t-1})(n_{t-1} - n_{t-2})\right] = \varphi E\left[(n_{t-1} - n_{t-2})^2\right] = \varphi\sigma^2\frac{(1-\varphi)^2}{1-\varphi^2}. \tag{9}$$

At lag k, the autocorrelation in NAV returns is φ^k. Formally, $\Delta n_t = (1-\varphi)(1-\varphi L)^{-1}r_t$, so the variance of NAV returns is $\sigma^2\frac{(1-\varphi)^2}{1-\varphi^2}$. Because φ is nonnegative, the autocorrelation in NAV returns is also nonnegative and increases with staleness. To the extent that there is stale pricing, NAV returns trend and are predictable. Similarly, the covariance of successive ETF price returns is

$$E\left[(p_t - p_{t-1})(p_{t-1} - p_{t-2})\right] = E\left[(u_t - u_{t-1})(u_{t-1} - u_{t-2})\right] < 0. \tag{10}$$

This follows because $\Delta p_t = r_t + \varepsilon_t + (\psi - 1)(1-\psi L)^{-1}L\varepsilon_t$. As the first two terms are innovations and the last term is negative, the cross product with the lagged price change is negative too. So, in contrast to NAV returns, the autocorrelation of ETF price returns is negative, reflecting the effect of transitory liquidity shocks that reverse over time.

4.2.4. Premiums. Using the definition of the premium $\pi_t = (p_t - n_t)$, we get

$$\pi_t = (v_t + u_t) - n_t. \tag{11}$$

With L the lag operator, i.e., $Ln_t = n_{t-1}$, then $n_t = (1-\varphi)(1-\varphi L)^{-1}v_t = (1-\varphi)(v_t + \varphi v_{t-1} + \ldots)$. Recall that from the AR(1) autoregressive formulation of liquidity shocks we have $u_t = \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \ldots$ so that the premium is $\pi_t = [1 - (1-\varphi)(1-\varphi L)^{-1}]v_t + u_t$.

Writing returns $r_t = (1 - L)v_t = v_t - v_{t-1}$ the premium can be noted as

$$\pi_t = \varphi(r_t + \varphi r_{t-1} + \ldots) + \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \ldots. \tag{12}$$

This shows the composition of the ETF's premium into two sets of terms: (*a*) price discovery (the product of the staleness factor and a weighted average of past fundamental returns) and (*b*) transitory liquidity (captured by a weighted average of past liquidity innovations). Several implications follow from this simple formulation. First, as the return and liquidity shocks have zero mean, the average premium mean reverts to zero over time. (For fixed income funds, the convention of using bid prices to compute the NAV implies a positive mean.) Second, a widening of the premium need not reflect greater liquidity demand but may reflect changes in fundamentals. Hasbrouck (2003) uses a vector error correction model to examine the information share of a particular security based on the relative contributions of that security's time series of innovations. He finds strong evidence that the S&P 500 ETF contributes to price discovery, especially for the sector's ETFs. It is important to understand this second point because some investors may avoid buying at a premium or selling at a discount when, in effect, the price of the ETF has moved to capture changes in value. Third, even if fundamental returns are serially uncorrelated, the premium still exhibits positive autocorrelation that increases with staleness and the slowness with which arbitrageurs correct pricing errors. Petajisto (2013) examines the deviation of midquote prices of ETFs from their NAVs, showing they are larger in funds holding international or illiquid securities. He controls for stale pricing using the cross section of prices for groups of similar ETFs and finds the average midquote pricing band is economically significant at 150 base points (see also Engle & Sarkar 2006).

4.3. Liquidity and Transaction Costs

Unlike open-ended funds, transactions can occur on the exchange throughout the trading day so that purchases/sales of ETFs do not necessarily require investors to interact directly with the fund. Although ETFs trade intraday on organized exchanges as equities, the unique creation-redemption mechanism allows the market to adjust the supply of available shares through primary market transactions in the underlying assets beyond the visible secondary market. This additional element of liquidity means that trading costs of ETFs are determined by the lower bound of execution costs in either the secondary or primary markets. Indeed, the bid-ask spreads of ETFs are frequently well below the corresponding costs of trading the underlying basket securities for both equities and bonds. Even in the absence of inventory costs, information asymmetry provides a compelling explanation for why ETF spreads would be much lower than in underlying basket securities.

The actual transaction price of the ETF is the midquote m_t plus or minus half the (effective) bid-ask spread at the time, denoted c_t, depending on whether flow is buyer or seller imitated:

$$p_t = m_t + \left(\frac{c_t}{2}\right) q_t. \tag{13}$$

Because quotations are two sided, it is reasonable in most cases to posit that the midquote reflects expected value, so $v_t = m_t$. The microstructure error term is then interpreted as $u_t = \left(\frac{c_t}{2}\right) q_t$. An exception would be if market makers could anticipate one-sided flow (e.g., during a sell off) and position their quotes accordingly, in which case the shock also captures this premium. The market maker's profit comes from simultaneously buying (selling) the ETF while offsetting or hedging the risk.

For an individual security, in the absence of other costs, the spread arises because order flow is informative and market makers protect themselves against adverse selection. The execution price is the conditional expectation given a purchase or sale: $E[v_t|q_t > 0] = m_t + \left(\frac{c_t}{2}\right)q_t$. In a portfolio context, the conditional expectation given that a broad index portfolio is being traded would weight only the common factor information component of flow and, hence, much lower spreads. For this reason—access to the underlying exposure at substantially reduced cost to acquiring that exposure directly—ETFs holding relatively less liquid portfolio securities often offer a compelling benefit to shareholders.

Figure 2 shows that the average time-weighted bid-ask spread (in basis points) for five ETFs with quite different exposures is significantly lower than the average spreads of the underlying securities, illustrating how ETFs can provide low-cost access to less liquid markets. This is especially so for bond funds where the over-the-counter nature of the underlying market (see, e.g., Hendershott & Madhavan 2014) can result in wide spreads. Large spreads in the underlying bonds often mean that the convention of pricing the NAV for bond funds using the bid price results in a positive premium.

4.4. Model Estimation

Empirical estimates of the model are of economic interest because they provide insights on the degree of staleness in the NAV, i.e., φ, the speed with which pricing errors are corrected, i.e., ψ, and the efficiency of the arbitrage mechanism measured by the estimated "true" deviations between price and (unobserved) value. The latter lets us decompose the premium to estimate the portion of the average premium attributable to liquidity versus price discovery. There are several ways to estimate the model.

First, note that the premium can be written as $\pi_t = \frac{\varphi}{1-\varphi}(n_t - n_{t-1}) + u_t$, where the error term u_t follows a first-order autoregressive process. An empirical formulation allows the premium to depend on other exogenous variables Z_t such as indicator variables for the crisis or short-selling restrictions, etc. We estimate

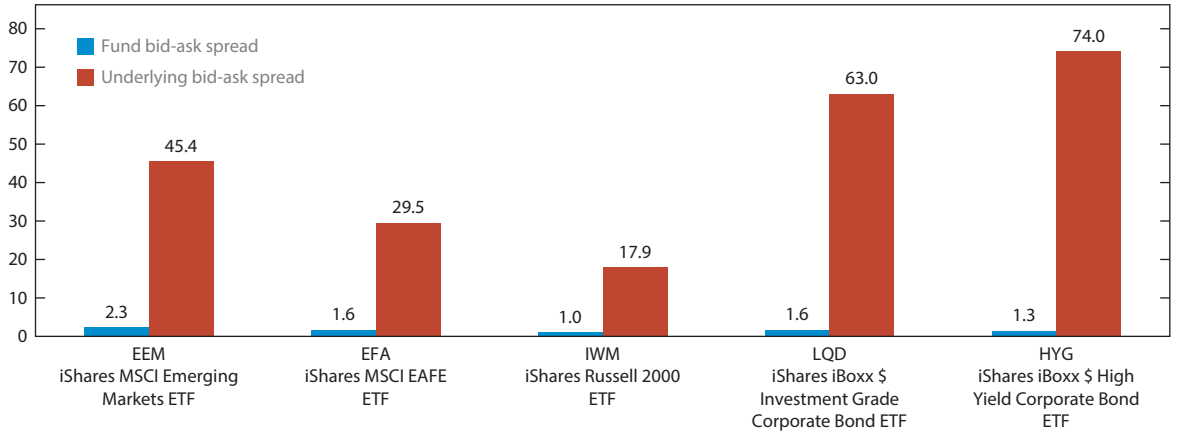

Figure 2

Bid-ask spreads (in basis points) of select ETFs and underlying baskets: (*blue*) fund bid-ask spread; (*red*) underlying bid-ask spread. Source: Bloomberg and TRACE data, June 2013. Abbreviation: ETF, exchange-traded fund.

$$\pi_t = \frac{\varphi}{1-\varphi}(n_t - n_{t-1}) + \gamma' Z_t + u_t, \tag{14}$$

which is well specified because both right- and left-hand sides are stationary. This regression model can be estimated directly with a generalized autoregressive error term to allow for higher-order effects. The estimated residuals from this model $\{\hat{u}_t\}$ then yield a time-series estimate of the "true" premium or discount. We can consistently estimate a measure of efficiency in terms of the estimated standard error of the residual, noting that $\sigma_u = \tau/\sqrt{(1-\psi^2)}$.

Second, the model may be directly and dynamically estimated by expressing it in a state-space form as follows. The measurement or observation equation is determined by price and the NAV, expressed as a function of the state (expected value):

$$\begin{bmatrix} p_t \\ n_t \end{bmatrix} = \begin{bmatrix} \psi p_{t-1} \\ \varphi n_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & -\psi \\ 1-\varphi & 0 \end{bmatrix} \begin{bmatrix} v_t \\ v_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}. \tag{15}$$

The transition (or state) equation is the unobserved expected value:

$$\begin{bmatrix} v_t \\ v_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_{t-1} \\ v_{t-2} \end{bmatrix} + \begin{bmatrix} r_t \\ 0 \end{bmatrix}. \tag{16}$$

This model can be estimated recursively using a Kalman filtering approach to estimate the true premium or u_t. Finally, the pricing relations or moment conditions can also be used to estimate the model using a generalized method of moments approach.

4.5. Performance and Benchmark Tracking

The terms "tracking error" and "tracking difference" are often used interchangeably, but they refer to different economic concepts. Tracking error is the ex post or realized volatility of the return difference between the ETF and benchmark portfolios, typically using daily returns, and measures the consistency with which an ETF follows its benchmark. The predicted tracking error is an ex ante concept based on the covariance matrix from a risk model. It is defined as the volatility given the difference between ETF and benchmark weights:

$$\sigma(h) = \sqrt{(h-h_b)' V (h-h_b)}. \tag{17}$$

Ex ante and ex post tracking error are helpful metrics for investors using ETFs for hedging purposes, but they do not provide an indication of how closely the longer-term performance of the ETF has matched its benchmark.

By contrast, tracking difference measures the actual under- or outperformance of a benchmark over a stated period of time. A physically based ETF may have high daily tracking error for institutional reasons (for example, different valuation sources or timing versus those used to calculate the benchmark), but it may still have a very low long-term tracking difference because the daily valuation differences net out over longer periods. In addition, standard calculations of tracking error ignore significant contributors to ETF performance, such as securities lending income. Tracking difference is the difference between the NAV return of the ETF and the benchmark return, which captures all income and costs that affect the ETF's NAV.

4.5.1. NAV versus index returns. Most index ETFs do not fully replicate their benchmarks in the sense that their holdings match completely with the components of the index. Instead, they will

hold a basket of securities designed to closely correlate with the return characteristics of the benchmark. They also may include a small percentage of "off-benchmark" holdings, such as cash, new issues, and futures. This is because many benchmark indexes comprise thousands of securities, including many that have small weightings and may be expensive or impossible to buy or sell. As a result, the performance of an ETF may not match that of its benchmark exactly, but it will effectively match the return of the specified market exposure.

The disparity in ETF and benchmark performance is known as performance deviation or, sometimes, tracking difference. (The term "tracking error" is also sometimes used but, as explained below, in many contexts that is an incorrect usage.) Some benchmark indexes are harder to match with correlated physical securities than are others, largely as a result of the number of investment factors that need to be correlated and the availability of index components that can be assembled to match those investment factors. Open-ended funds and other index-tracking investment vehicles face similar issues, but ETFs are more commonly offered on some of the harder-to-match benchmark indexes.

> **Example 2: Index Replication** In some cases, legal and regulatory constraints may prohibit an ETF from literally replicating its index. For example, one stock may represent 40% of an index's weight, although US ETFs are subject to a tax diversification rule that prohibits the largest holding from exceeding 25% of assets and requires that holdings greater than 5% of assets be limited to a collective sum below 50%. European ETFs are subject to a UCITS (Undertakings for Collective Investment in Transferable Securities) rule that prevents an ETF from investing more than 5% of its assets in securities issued by a single issuer. ETFs are sometimes managed against "capped" indexes (where available), which limit the maximum weights of the largest holdings. Finally, some ETFs may exhibit significant performance deviation from their benchmark indexes because the ETF manager makes choices that allow for greater tracking error in an effort to achieve other goals, such as enhanced liquidity of the ETF's shares. Illiquid securities included in a benchmark index but excluded from an ETF's portfolio may also not perform in tandem with other components of the index: For example, if the ETF is systematically underweighted to the differently performing illiquid securities, the ETF's performance will reflect its liquidity bias.

4.5.2. Beta solutions. Futures and ETFs are vehicles for index exposure, but they offer different trade-offs. Index futures and ETFs are used for a variety of investment purposes, including

- Investing excess cash
- Hedging exposures
- Shifting synthetic exposure rapidly
- Implementing long-term strategic risk allocation.

A quantification of the trade-offs between the two beta solutions requires an understanding of their true costs/benefits and pricing. Trade-offs change dynamically with market conditions (e.g., securities lending revenues, dividend yields, roll costs) and investor objectives (turnover, tracking error). In recent years, the range of exposures attainable through ETFs has greatly increased, their ownership costs have decreased, and their liquidity has generally improved, whereas structural factors (e.g., regulation and broker capital) have increased costs for futures.

Whereas ETF pricing reflects the value of the underlying basket plus a transitory premium or discount, as described in Section 4.2, futures pricing is more complex. The cost of buying and

carrying a portfolio of underlying securities in the index is the "cost of carry," which measures the difference between spot S_t and futures prices over the interval $F_{t,T}$ as a function of the financing rate r less the dividend yield d. Assuming nonstochastic interest rates for expositional simplicity, the futures price (in this case also the forward price) is

$$F_{t,T} = S_t e^{(r-d)(T-t)}. \tag{18}$$

Rolling a futures position from the nearby contract to the deferred contract need not accurately capture the movements in the spot price. Comparing the financing cost to the implied financing rate determines roll "cheapness" or "richness." The cost of carrying a position in the underlying equities portfolio as implied by futures prices may deviate from its true economic cost ("fair value"). The cost of carry uncertainty leads to pricing risk, and changes in cost of carry assumptions may lead to profits or losses.

Example 3: E-Mini S&P 500 and MSCI Emerging Market Futures As an example, the E-Mini S&P 500 and MSCI Emerging Market futures in the June 2013 roll averaged on an annualized basis LIBOR + 0.37% and 0.45%, respectively, as illustrated in **Table 3**. These figures are consistent with recent roll patterns. For example, if 3-month LIBOR was 0.30% and the current roll richness was L + 0.37%, the cost of owning S&P 500 futures for 1 year would be approximately 0.67% versus a typical ETF management fee from 0.07% to 0.10%, indicating lower costs for ETF investors seeking exposure to the S&P 500 index.

Table 3 Roll volumes and costs as a premium to LIBOR for select futures contracts

	S&P 500 E-Mini futures		MSCI Emerging Markets futures	
Days to expiration	Volume	Implied financing rate	Volume	Implied financing rate
14	23,119	0.71%	3	0.55%
13	24,487	0.70%	1	0.59%
12	27,670	0.70%	115	0.79%
11	39,343	0.67%	416	0.83%
10	38,926	0.67%	391	0.67%
9	109,891	0.69%	3,086	0.78%
8	150,169	0.70%	3,568	0.77%
7	326,228	0.67%	9,305	0.87%
6	334,656	0.67%	24,041	0.85%
5	459,289	0.66%	45,460	0.89%
4	491,896	0.66%	47,714	0.88%
3	535,531	0.65%	53,299	0.76%
2	360,778	0.67%	31,109	0.61%
1	388,392	0.57%	28,311	NA

Source: CME Group and NYSE LIFFE. Roll data as of June 2013 for roll period prior to expiration of June 2013 contracts.

Futures can trade "rich" or "cheap" to fair value when futures curves are in either contango or backwardation. Roll costs are not solely dependent on the term structure of the futures curve; they also reflect liquidity pressures or one-sided markets. In particular, there is evidence that long investors roll exposures at roughly the same time, causing departures from fair value.

In practice, many different return drivers may lead to performance differentials between ETFs and futures (as illustrated in **Table 4**), including

- Cost of carry drivers: implied financing rate of equity exposure, dividends/income, return on cash invested from lack of futures capital commitment, management fees
- Implementation cost: transaction costs from executing ETF position or rolling/maintaining continuous futures position.

Morillo et al. (2012) note that in the period June 2000 to September 2011, an ETF implementation averaged −2 basis points per quarter of slippage against the S&P 500 index versus −13 basis points with fully funded futures contracts.

5. ALTERNATIVE BETA ACTIVE STRATEGY

The great majority of ETFs passively track an index, but an increasing number pursue active strategies of various forms, as depicted in **Table 1**. For our purposes, we define any deviation from market capitalization weighting as an active strategy, with the logic deriving from asset pricing theory where the market portfolio is mean-variance efficient. Here we consider first deviations from market capitalization weights based on considerations other than alpha (e.g., risk or liquidity considerations) and then turn to alpha strategies that directly attempt to outperform a benchmark. Here, alpha refers to the excess fund return performance relative to its stated benchmark, and the term alternative beta refers to weighting schemes, either heuristic or based on formal optimization, that deviate from market capitalization weighting. In some cases, the motivation for alternative beta is a concern that market capitalization weights are too concentrated in larger cap names or, in the case of fixed income products, provide too much exposure to issuers who are most in need of borrowing.

Table 4 Key elements of return differentials in beta exposure vehicles

Return sources	Exchange-traded funds	Futures
Asset price return	Underlying basket return	Index return
Dividends	+ Actual dividends	+ Forecasted dividends
Securities lending	+ Securities lending income	+ Forecasted securities lending income
Financing rate	None	− Forecasted financing rate
Cash yield	None	+ Actual cash yield
Management fees	− Management fees	None
Transaction costs	− Cost of purchase and sale roundtrip	− Cost of purchase and sale roundtrip − Cost on futures roll − Cost on cash management
Changes in premium/discount to fair value	Change in premium (+) or discount (−)	Change in premium (+) or discount (−)

5.1. Alternative Weighting Schemes and Fundamental Indexes

The simplest example of nonmarket capitalization weights is equal weighting, where the portfolio weight on stock $i = 1, \ldots N$ is $h_i = 1/N$. A related but more complex heuristic approach is so-called diversity weighting, which is mathematically a blend of capitalization- and equal-weighting schemes. Note that price-weighting schemes (where $h_i = p_i/\sum_k p_k$) have a long history, including the Dow Jones Industrial Average and Nikkei 225. Equally weighted portfolios typically tilt toward the size factor, in that they emphasize smaller capitalization securities and have greater portfolio turnover than an equivalent capitalization-weighted benchmark.

So-called fundamental indexes (see, e.g., Arnott, Hsu & Moore 2005) overweight securities with characteristics sought by investors, including the following:

- Firm size: revenues and book value
- Profitability: profits, margins, and cash flow
- Risk: idiosyncratic and systematic
- Dividends: yield, consistency, and sustainability.

Once the benchmark has been selected (which in and of itself is an active decision), the fund replicates the holdings to the extent possible. The stated motivation for fundamental indexes is to avoid errors induced from market capitalization weighting—the idea is that a positive (negative) pricing error in a stock leads to over- (under)weighting in the index. As these errors correct over time, proponents argue, a market capitalization–based weighting scheme will exhibit return drag. However, this argument implies that market capitalization should be a predictor of future excess returns (see Perold 2007). Thus, fundamental indexing is a form of active management because it involves a tilt of some sort based on alpha drivers such as earnings, etc.

5.2. Model-Based Transparent Active Exchange-Traded Funds

Model-based solutions add a variety of enhancements to the heuristic approaches to risk reduction including control of turnover, constraints on tracking error, or deviations from sector/country weights. An important category of alternative beta is the class of model-based transparent active where explicit quantitative rules govern portfolio selection. Transparent in this regard means the fund's holdings are visible on a daily basis with a one-day lag. These include minimum-variance portfolios, which minimize risk subject to a fully invested constraint; factor portfolios, which seek exposure to various economic factors; and various diversification schemes. Here I develop a general framework to accommodate various types of active model-based strategies to select optimal holdings. The inputs are an N vector of alphas α and an $N \times N$ covariance matrix V. In the minimum-variance portfolio, $\alpha = 0$. By contrast, for factors or other active portfolios, α takes on positive or negative values on the basis of the quantitative or model-based signal.

Consider a generalized optimization problem that selects the optimal active weight h to maximize portfolio alpha $h'\alpha$ against active risk $h'Vh$ (weighted by a risk-aversion parameter ρ), subject to the constraint that the portfolio holdings add to 1 (i.e., are fully invested). Note that, for simplicity here, the holdings are relative to the benchmark, which could be interpreted as market capitalization–based weights on the desired index. Formally, we write

$$Max\ U(h) = h'\alpha - \rho h'Vh, \tag{19}$$

subject to $(h'e - 1)$, where $e = (1, \ldots, 1)$ is an $N \times 1$ vector. Maximization of the objective function over h with the additional assumption that the sum of all alpha over- and underweights equals zero yields the optimal portfolio:

$$h^* = \left(\frac{1}{2\rho}\right) V^{-1}\alpha + \left(\frac{V^{-1}e}{e'V^{-1}e}\right). \tag{20}$$

Without the additional constraint of full investment, unrestricted active holdings are

$$h^* = \left(\frac{1}{2\rho}\right) V^{-1}\alpha. \tag{21}$$

So, the strategic beta portfolio consists of two terms: The first term is the active portfolio, based on alpha forecasts, and the second term, as discussed below, is the minimum-variance portfolio. The requirement to be fully invested is a key differentiator from a classic long-short equity fund.

5.3. Minimum Volatility Strategies

Minimum-variance strategies have long been attractive (see, e.g., Black 1972) because of their high Sharpe ratios over time, low correlation with individual security selection models, and consistency across markets. Sensibility arguments include either behavioral explanations or technical mispricing of risk where low-risk stocks have abnormally high risk-adjusted returns. Behavioral hypotheses include investor attention or risk-seeking behavior where high-risk, high-beta stocks are more widely followed and are overvalued. Technical explanations include constraint-based theories where investors are unable to fully gain their desired leverage and, hence, pay a premium for high-beta stocks. Setting $\alpha = 0$ in the expression for h^*, we obtain the pure minimum-variance portfolio, which minimizes total portfolio risk while requiring the holdings to sum to 1:

$$h^* = \left(\frac{V^{-1}e}{e'V^{-1}e}\right). \tag{22}$$

Clarke, de Silva & Thorley (2006) provide empirical evidence on the performance of minimum-variance portfolios relative to their market capitalization–weighted analogs.

5.4. Model-Based Diversification Strategies

Other model-based approaches also seek to reduce portfolio risk or concentration in ways that resemble minimum-variance portfolios. Here, risk parity and maximum Sharpe objective functions are discussed. Risk parity achieves diversification by finding weights that equalize the contributions of each security to portfolio risk. Formally, we find portfolio weights for every security i and j such that

$$h_i \frac{\partial\sigma(h)}{\partial h_i} = h_j \frac{\partial\sigma(h)}{\partial h_j}. \tag{23}$$

As with the minimum-variance portfolio, this weighting scheme depends only on the risk model through the estimated variance matrix.

Another objective is to maximize the portfolio's Sharpe ratio, i.e., find the portfolio with the highest ratio of ex ante return (in excess of the risk-free rate) to volatility. Unlike the methods above, this objective requires assumptions on expected returns. One approach is to assume the expected excess return for every stock $i = 1, \ldots N$ is proportional to its own volatility so that $E[r_i] - r_f = \rho\sigma_i$. In this case, the maximum Sharpe ratio portfolio requires selecting h to

maximize $h'\hat{\sigma}/\sqrt{h'Vh}$ subject to the constraint of full investment ($h'e - 1$). In the numerator, $\hat{\sigma}$ is the $N \times 1$ vector of stock volatilities σ_i defined as the square root of the diagonal elements of the variance matrix V. Accordingly, the maximum Sharpe portfolio also depends only on the assumed risk structure; in practical terms, this closely resembles the approaches described above.

5.5. Optimized Funds

Optimized funds that balance volatility against other factors such as transaction costs/liquidity offer another example where $\alpha = 0$ but the fund departs from market capitalization weighting. Letting $c(h_t - h_{t-1})$ denote the transaction costs associated with trading (i.e., with the change in holdings), the objective function would be to choose h to minimize a weighted sum of variance and illiquidity:

$$Min_{\{h\}} U(h_t) = (h_t - h_b)'V(h_t - h_b) + \gamma c(h_t - h_{t-1}), \tag{24}$$

where γ is the penalty on transaction cost and turnover. The transaction cost function could be simply a measure of liquidity such as bid-ask spreads in the underlying securities, in which case the functional form is linear and c is half the bid-ask spread. In a more complex model with market or price impact, the functional form would capture the costs associated with a change in holdings, thereby introducing a time dependency.

5.6. Factors

Recently, factor portfolios have also become popular for investors seeking exposure to particular factors such as value or momentum. Academic and industry research has identified a number of factors that, over long periods of time and across regions, have outperformed broad market indexes. These include the following:

- Size: small cap has traditionally outperformed large cap stocks
- Value: measured by metrics such as book/price, earnings/price
- Quality: earnings quality and variability; lower leverage
- Momentum: positive return momentum.

As with fundamental indexes, factor investing seeks to capture longer-term risk premia, and the tilts from cap weighting are typically very similar. Our framework can be used to assess factor portfolios. Observe that the alpha that derives from a K-factor model is not the sum of the alphas from each of the individual factors. That is, if we have a model where alpha is comprised of signals $k = 1, 2, \ldots, K$ with weights w_k, the optimal holding vector for each individual factor is

$$h_k^* = \left(\frac{1}{2\rho}\right) V^{-1}\alpha_k + \left(\frac{V^{-1}e}{e'V^{-1}e}\right). \tag{25}$$

Then, $h^* = \sum_{k=1}^{K} w_k h_k^*$ does not generally occur because the covariance matrix weights the active portion of the portfolio. In other words, an optimal portfolio with exposure to multiple factors is not the same as a weighted average of single factor portfolios. Intuitively, the stocks in each portfolio are weighted differently, and the constraint of being fully invested does not allow offsets in weights that are possible when building a single, optimized portfolio.

6. FUNDAMENTAL ACTIVE EXCHANGE-TRADED FUNDS

6.1. Reverse Engineering

One concern with active ETFs that are following proprietary strategies is the ability of others to infer the underlying alpha or excess return forecast. Unlike most open-end and closed-end funds, ETFs disclose all (or substantially all) portfolio holdings on a contemporary basis to facilitate secondary market trading. Portfolio transparency raises the possibility of revealing information about changes in an ETF's portfolio that may be market sensitive and raises the cost of executing future transactions. Although not generally a concern for model-based strategies of the types discussed above, this may make it difficult to pursue certain active investment strategies (e.g., security selection in small-capitalization stocks or thinly traded bonds) through ETFs. Because holdings are published daily, the reported holdings h_t can, in theory, be inverted to estimate alpha:

$$\hat{\alpha}_t = 2\rho V\left[h_t - \left(\frac{V^{-1}e}{e'V^{-1}e}\right)\right]. \tag{26}$$

So, α can be recovered up to a scalar, because the fund manager's risk-aversion parameter ρ is not observable to an outsider. This exercise presumes that covariance matrix V is observed, which is a reasonable assumption given that well-established risk models yield similar estimates of variances and covariances.

With a time series of holdings, the signals underlying the alpha forecast can, in theory, be reverse engineered, leading to a conundrum: If the manager can successfully forecast excess returns, then their signals can be readily copied if implemented in an ETF structure, which by nature is highly transparent. Successful managers will then find their intellectual property quickly eroded. Some ETFs do use published hedge fund holdings (through 13-F filings) to create portfolios of successful fund managers. These funds operate on the assumption that published holdings, which are available only with a substantial delay, are representative of current positions and, more importantly, that the hedge fund managers they are mimicking are consistent in their generation of alpha. As positions may be changed quickly, there is less opportunity for this type of mimicry to erode alpha returns over time.

A closely related concern with active funds is front running. Because alpha is based on a set of underlying signals (e.g., book/market, earnings/price, quality, etc.) that will typically decay slowly,

$$\alpha_t = \gamma\alpha_{t-1} + \xi_t, \tag{27}$$

where $0 < \gamma < 1$ captures the autocorrelation in alpha and ξ_t is a shock that captures new signal innovations. Thus, if holdings are publicized, future trades can theoretically be forecast, especially if the manager uses a transaction cost function to scale trades to avoid market impact, leading to a dependence of trades (i.e., $h_t - h_{t-1}$) on current and past alphas. Trading ahead could increase the fund's transaction costs and erode its alpha.

6.2. Nontransparent Active Exchange-Traded Funds

An alternative in theory is a more opaque or nontransparent structure where exact holdings are not public (e.g., the manager provides general portfolio characteristics) or are not reported in a timely manner. A concern for ETFs, unlike active mutual funds, is that the creation-redemption mechanism needs transparency to keep bid-ask spreads and tracking error low. However, there may be effective ways to convey the risk characteristics of the fund (much as with a principal bid) for hedging purposes without divulging exact positions or by using a blind-trust mechanism.

Nontransparent active funds based on fundamental (nonquantitative) styles of security selection or opportunity seeking (i.e., trading on availability, buying assets at the bid, and selling at the ask) present another opportunity for investors and are difficult to mimic.

Even if holdings are not published or are not published regularly, as in a nontransparent ETF structure, the reporting of intraday NAV means that holdings can be approximated. For example, suppose NAV is reported at intervals of $\Delta t = 15$ s, as is current US practice. This means that we observe a value equal to $n_t = \sum_{i=1}^{N} m_t^i h_i$ every 15 s, where m_t^i is the last recorded midquote of asset i at second t and h_i is the active holding. Market participants observe midquotes (whether stale or not), but the underlying holdings are not disclosed. Each observation of intraday NAV is thus a linear equation in N unknowns, meaning the active weights h_i. To estimate the N unknown holdings vector (assuming positions are constant) requires at least N equations of this type, i.e., a calendar duration of $T = \frac{N}{\Delta t}$. Estimation is straightforward as the NAV is a linear function. Let M denote the $N \times N$ matrix of observed midquote prices (row k of this matrix contains the midquotes for each of the N assets at period k) and y the $N \times 1$ vector of reported intraday NAVs. Then, assuming M has full rank, the estimated average holding is

$$\hat{h} = M^{-1} y. \tag{28}$$

For example, if the universe consists of 500 assets, approximately 2 h and 5 min are needed to solve for the manager's active weights. With longer durations beyond the minimum needed of $T = \frac{N}{\Delta t}$, this computation can be rolled forward over the day or it can be used to overidentify $\hat{h}$ and provide a statistical confidence interval using the generalized method of moments approach.

As a practical reality, however, concerns regarding reverse engineering and front running are unwarranted. First, managers may change holdings during the day so any reverse engineering is only an approximation to the average active weight over that interval of time. Furthermore, intraday NAV is typically computed against yesterday's holdings, so in the example above, estimations on yesterday's average holding are being made on a rolling 2-h basis. Note also that intraday NAVs for certain illiquid or international funds can be quite misleading. Second, portfolio managers may employ multiple implicit or explicit constraints (e.g., sector or industry caps, etc.), and the resulting portfolio becomes very difficult to reverse engineer owing to the nonlinear impact of constraints. Third, even if the portfolio were reverse engineered and replicated (presumably with less tax efficiency), the investors in the fund would not be impacted except to the extent that their alpha is eroded. Finally, note that mutual funds following active strategies are no different in this regard.

7. IMPACT OF EXCHANGE-TRADED FUND FLOWS

There is a popular perception that the growth in passive index investing through ETFs has had detrimental effects on the market quality of the underlying basket securities. In particular, there is concern that ETF trading substitutes for and takes away from volume and liquidity in the underlying securities and increases the comovement in their returns. In turn, it is argued that increased pairwise return correlation impairs price discovery and the ability of active managers to generate alpha. For instance, Wimbish (2013) cites concerns that index-focused ETFs cause greater stock return correlations and contribute to systemic risk.

Let us define the signed flow into a particular ETF on day t as total creations less redemptions in value terms or the change in shares outstanding times the NAV of the ETF:

$$f_t = (o_t - o_{t-1})n_t. \tag{29}$$

Flows are not the sum of signed intraday volumes (where buyer- or seller-initiated volumes are classified as those executed at the ask or bid prices, respectively, and then buy volumes less sell volumes are added) because intraday ETF trades may net out without a creation or redemption.

Empirical analyses of the impact of passive flows are relatively scarce. A notable exception is Sullivan & Xiong (2012), who argue that index investing contributes to higher cross-sectional trading commonality and higher return correlations among stocks. Da & Shive (2013) find a strong relation between measures of ETF activity and return comovement at both the fund and the stock levels. As a result, they conclude that some ETF-related return comovement is excessive and that ETFs reduce the diversification benefits they were intended to promote.

In turn, increased comovements in the returns of ETF basket securities are alleged to have had a variety of negative impacts with themes including reduced diversification benefits, greater difficulty in stock-picking during periods of high correlation, and the impairment of price discovery and reduction in liquidity in smaller capitalization stocks. Flood (2012) further notes that active managers face headwinds in the form of an increase in correlations between constituents of the Russell 2000 index which coincided with the growth of small cap ETFs as pairwise correlations for stocks in that index rose from 5% in 2000 to 25% in 2012. Similarly, the fact that ETFs were disproportionately affected in the Flash Crash of May 6, 2010 (ETFs accounted for 70% of trades ultimately cancelled by exchanges), has fueled discussion regarding ETF flows, return volatility, and systemic risk (see Bradley & Litan 2010; Ramaswamy 2010; Wurgler 2010; Ben-David, Franzoni & Moussawi 2011). The role of leveraged ETFs has also been discussed (see, e.g., Cheng & Madhavan 2009) in the context of end-of-day volatility effects.

7.1. Impact of Flows on Underlying Constituent Returns

Along with the growth of ETP assets, cross-stock return correlations have increased sharply over the past decade, leading some commentators to draw a causal connection. The idea is intuitively appealing: ETF flows cause stock prices in the underlying baskets to move more together. But does this make sense conceptually or empirically? Madhavan & Morillo (2014) argue that it does not. First, arguments attributing the higher correlation environment to ETF growth frame the problem incorrectly. If ETFs were not a viable investment vehicle, investors seeking broad-based exposures have many other alternatives such as active mutual funds. Second, pairwise correlation is ultimately a secondary statistic, driven by the ratio of factor volatility to specific volatility. Common factor effects ultimately drive changes in pairwise correlation because significant changes in idiosyncratic risk are unlikely. Third, this argument, however, is not convincing in the longer historical and macroeconomic context. Indeed, even though cross-stock correlations are high today, they are not at unprecedented levels relative to the past, well before the rise of passive indexing. **Figure 3** contains two time series from January 1927 to December 2012: the annualized volatility based on daily equity returns, computed month by month, and the corresponding monthly correlation among the stocks based on daily returns in the month. The data source is CRSP (Center for Research and Security Prices) for the largest 200 stocks prior to 1958 and S&P 500 thereafter.

Two points are immediate: (*a*) Correlations were very high in the past, well before the growth of index and ETF investing, and (*b*) correlations and volatility are clearly related—macrouncertainty

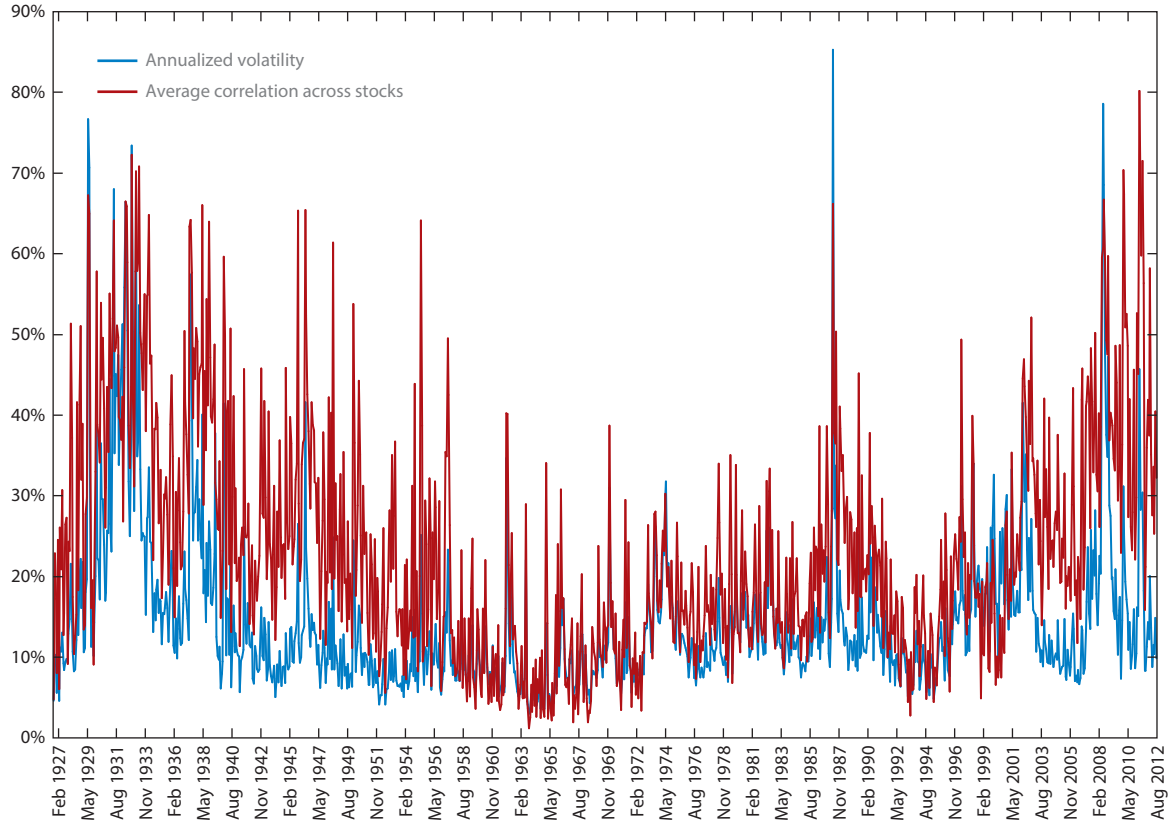

Figure 3

Time series of active and passive funds from January 1927 to December 2012: (*blue*) annualized volatility; (*red*) average correlation across stocks.

drives overall factor volatility and, hence, correlation. Furthermore, asset classes such as currencies also show an increase in correlation despite limited ETF penetration. This is consistent with the hypothesis that the driver of higher correlations is the macroenvironment and not growth in ETFs (also see Mazza 2012).

In terms of the purported linkage between stock selection ability and the correlation environment, there is no logical connection. Indeed, at any correlation level, active managers (those who deviate from the cap-weighted distribution of holdings in the universe) must, by definition, have average performances equal to the benchmark return less fees. Thus, correlation is not linked to the success of active management and stock selection in particular.

Recent discussion has taken the view that trading in the ETF, possibly quite transitory in nature, leads to a propagation of volatility into the underlying securities. For example, Broman (2013) argues, "ETF mispricing" is only partially mean reverting. Ben-David, Franzoni & Moussawi (2011) posit that ETF flows transmit liquidity shocks to their underlying baskets of securities. In their view, creation/redemption activity adds a layer of volatility to the underlying security returns as opposed to the model presented here in which arbitrage gaps are self-correcting. Their empirical evidence is based on regressions of NAV returns on the lagged premium, lagged NAV return, and lagged ETF return, or assuming all prices and values are in logs,

$$(n_t - n_{t-1}) = \beta_0 + \beta_1 \pi_{t-1} + \beta_2 (n_{t-1} - n_{t-2}) + \beta_3 (p_{t-1} - p_{t-2}) + \varepsilon_t. \quad (30)$$

The coefficient on lagged ETF returns is interpreted as shock propagation. Given the decomposition of the premium derived earlier, the left- and right-hand-side terms both include weighted sums of past returns so the regression estimates are difficult to interpret. In other words, as shown above, the alternative explanation is that price impact reflects shocks to fundamentals, ETFs lead price discovery, and the NAV is "stale" and "catches up" over time. The stylized facts from this regression can be interpreted either way, with no evidence of causality pointing in either direction.

7.2. Leveraged and Inverse Exchange-Traded Funds

Leveraged and inverse exchange-traded products (LETPs) provide leveraged long or short exposure to the daily returns of various indexes, sectors, and asset classes. As these products are not physically based, the term ETP, as opposed to the more conventional ETF, is used. The LETP space has attracted significant assets and now comprises leveraged, inverse, and leveraged inverse products offering long exposure of 2× or 3× or short exposure of −1×, −2×, or −3× the underlying daily index return. Several factors explain the attraction of LETPs. First, these products offer short-term traders and hedge funds a structured vehicle to express their directional views regarding a wide variety of equity indexes and sectors. Second, unlike traditional ETFs, these products have leverage embedded in their design, so investors need not explicitly use derivatives or margin. Third, investors—attracted by the convenience and limited liability nature of these products—increasingly use them to place longer-term leveraged bets or to hedge their portfolios. Despite their popularity, leveraged products have been controversial for several reasons discussed below.

7.2.1. Compounding and returns. A concern with LETPs is that some investors may not recognize the impact of the compounding of daily leveraged returns over longer intervals, as illustrated below.

> **Example 4: Compounding of Levered Returns** Consider a 2× LETP with an initial NAV of $100. Suppose the underlying index falls 10% on day one from its initial value of 100 to 90 and then goes up 10% on the subsequent day to 99, for a two-day decline of −1%. Although investors may expect the leveraged product to decline by twice as much, or −2%, over the two-day period (as shown above, interest rates are also a factor), it actually declines by −4%. This is because doubling the index's 10% fall on the first day lowers the LETP's value from $100 to $80, which then recovers to $96 (= 2 × 10% × $80) the following day, upon doubling the index's 10% gain.

Although simple, Example 4 provides valuable intuition. However, it is useful to model the underlying return process more formally to better understand return dynamics. Cheng & Madhavan (2009) develop a continuous time model where the benchmark index St follows geometric Brownian motion with a nonstochastic drift rate μ and volatility σ. Denoted by W_t, a Wiener process with a mean of zero and a variance of t,

$$dSt = \mu St\, dt + \sigma St dW_t. \quad (31)$$

As time intervals shrink and microstructure frictions are ignored (i.e., $u_t = 0$), the instantaneous return on an LETP with leverage x over an interval $[t, t + dt]$ is given by

$$\frac{dpt}{pt} = x\frac{dSt}{St} - (x-1)r_f dt, \tag{32}$$

where r_f is the risk-free rate (see Jarrow 2010, who provides a generalized construction of this model with nonzero risk-free rates that characterizes the return distribution of the LETF over any investment horizon). Thus, the instantaneous return on the LETF is x times the return on the index, less the interest factor. Under the assumptions of nonstochastic drift and variance provided by Cheng & Madhavan (2009), and with constant interest rates, the LETP's total return from time 0 to T, denoted $r_{0,T}$, can be written as

$$r_{0,T} = \left(\frac{S_T}{S_0}\right)^x e^{-(x-1)T\left(r_f + x\sigma^2/2\right)}. \tag{33}$$

Observe that the LETP's total return is the product of x-levered index returns, i.e., $\left(\frac{S_T}{S_0}\right)^x$, multiplied by a term that is less than 1. Accordingly, a buy-and-hold investor experiences a convexity reduction that increases with volatility, leverage, interest rates, and the time horizon. The model is straightforward to extend to incorporate complexities such as dividends, stochastic volatility, jump diffusions, and fund expenses, but our fundamental insights are not altered.

Giese (2010) shows that the results of Cheng & Madhavan (2009) hold under generalizations for multiasset portfolios with stochastic volatility. He shows there is an optimal degree of leverage that maximizes the expected future return of the daily rebalanced leveraged investment strategy that depends on observable market parameters. Further insight is provided by Haugh (2011), who shows that an investor with a long position in an LETP is short realized variance and interprets the exposure of a nonleveraged constant-proportion strategy to realized variance as a multiplicative premium. Jarrow (2010) provides a construction with an LETF and riskless bond that characterizes the return distribution of the leveraged ETF over any investment horizon. Other concerns for longer-term investors in LETPs include tax efficiency and transaction costs arising from higher turnover. Avellaneda & Zhang (2009) and Lu, Wang & Zhang (2009) empirically examine how closely LETPs correlate to reproducing the corresponding multiple of index returns over extended investment horizons.

> **Example 5: Return Deviations** A real-world example is the ProShares Ultra Short Oil & Gas −2× inverse product. Between August 2 and November 15, 2011, the underlying Dow Jones US Oil & Gas Index lost 3.9%, but the actual ETP experienced a loss of −9.9%.

7.2.2. Mechanics of rebalancing. LETPs generally rely on total return swaps to produce returns that are a multiple of the underlying daily index returns. The exposures of total return swaps underpinning LETPs must be rebalanced daily to produce the leveraged returns. Cheng & Madhavan (2009) show that LETPs (including inverse funds) induce rebalancing activity toward the end of the day in the same direction of the market, which is, in effect, similar to same-direction trading induced by portfolio insurance in the 1980s. When the underlying index is up, additional total return swaps exposure must be added. However, when the underlying index is down, the exposure of total return swaps must be reduced. This is always true whether the products are leveraged, inverse, or leveraged inverse.

Indeed, the effect of inverse products is very nonlinear. Specifically, Madhavan & Cheng (2009) show that the estimated end-of-day rebalancing (aside from creation/redemption activity) denoted by B_t can be expressed as the product of three terms: the assets under management at the start of the day, a function of leverage, and the index return for the day. With a fund with leverage x, this expression is

$$B_t = o_t n_{t-1}(x^2 - x)r_t. \quad (34)$$

Because the function of leverage in Equation 34 is $(x^2 - x)$, the effects of rebalancing are highly nonlinear. For example, for a triple-leveraged ETF $x = 3$ the function equals 6 $(= 3^2 - 3)$, whereas for a triple-inverse fund $(x = -3)$ it is 12 $[= (-3)^2 - (-3)]$.

> **Example 6: Leverage Rebalancing** Consider a $-2\times$ LETP with an initial NAV of \$100. At the outset, the required notional amount of the total return swaps is $-\$200$ (i.e., short twice the NAV). If the index falls 10% on day one, the fund's NAV increases 20% to \$120, and the exposure of the total return swaps goes to $-\$180$, reflecting the initial short exposure plus the \$20 gain from the day. Now, the required notional amount for the total return swaps is $-\$240$, or $-2\times \$120$. So, the fund will need to increase its short exposure of total return swaps. The change in exposure is $-\$60$ $(= \$180 - \$240)$, which the swap counterparty will presumably need to short in turn to avoid undue market risk. We can verify the formula given in Cheng & Madhavan (2009) that the implied rebalance is assets under management $\times$ index return $\times$ leverage factor, or $-\$60 = \$100 \times (-10\%) \times [(-2)^2 - (-2)]$.

Why is the direction of the effect the same for LETPs that are short the index? Intuitively, an inverse or leveraged inverse product's NAV will increase if the index falls, which requires it to increase short exposure still further, generating selling pressure. In other words, there is no offset or "pairing off " of leveraged long and short products on the same index. Note that the need for daily rebalancing is unique to LETPs owing to their product design. Traditional ETFs that are not leveraged or inverse (whether they are holding physicals, total return swaps, or other derivatives) do not induce such daily rebalance activity.

7.2.3. Impact of rebalancing activity. In theory, the rebalancing activity of LETPs should be executed as near the market close as possible, given the dependence of the rebalancing amount on the close-to-close return of the underlying index. Whether LETPs rebalance their exposure of total return swaps immediately before or after the market close, however, the counterparties with which they execute total return swaps will want to put on or adjust their hedges while the market is still open, to minimize the risk to their capital and position taking, especially in volatile markets. As LETPs gain assets, there may be a heightened impact on the liquidity and volatility of the underlying index and the securities comprising the index during the closing period of the day's trading session (e.g., the last hour or half-hour). As shown above, the rebalancing impact of increased assets is magnified by an increase in leverage and a shift in the asset mix toward inverse products.

Furthermore, as rebalancing flows are always in the same direction as the market movement, LETPs cannot by themselves mechanically cause whipsawing in the market, as the rebalancing induced by their need to keep to their leverage targets is, by construction, in the same direction as the market. Trading in the same direction as the market can, however, result in higher volatility, simply by increasing the overall scale of market moves as a result of the price impact that rebalancing may have into the close.[3] Whether this potential effect is material is ultimately an empirical question that depends on the size of the rebalances (and, in turn, is dependent on the size and

[3]Note that the use of leverage is not the key issue, as investors had access to instruments that allow for taking leveraged positions across a wide range of assets well before LETPs were available. The key issue is associated with the fact that LETPs reset their leverage on a daily basis and, thus, have to rebalance mechanically toward the close of the market, something that may not otherwise be the optimal implementation choice for some investors, particularly those interested in long-term positions.

leverage of these products), the proportion of those rebalances that are actually traded in the market on any one day, and the price impact associated with those trades. Recent evidence is presented by Tuzun (2013), who finds that price-insensitive and concentrated trading of LETPs could be destabilizing during periods of high volatility, especially for the stocks of financial firms (see also Haryanto et al. 2013, who conclude the impact was economically significant only during volatile periods). From a financial stability viewpoint, LETPs rebalancing could amplify a large stock market move and trigger a cascade effect through further rebalancing.

7.3. Flash Crash

The "Flash Crash" of the afternoon of May 6, 2010, saw the Dow Jones Industrial Average drop almost 1,000 points in 20 min. Multiple securities traded at clearly unreasonable prices, including some well-known stocks that traded at pennies. Notable was the disproportionate representation of ETPs among the securities most affected, with prices diverging widely from their underlying NAVs. Despite its short duration, the Flash Crash affected many market participants. Exchanges ultimately cancelled trades at prices below 60% of the 2:40 PM exchange-trade price, but many retail investors with market stop loss orders still had orders executed at prices well below prevailing market levels earlier in the day. Professionals who bought at distressed prices and hedged by short-selling similar securities or futures contracts incurred steep losses as their long positions were ultimately cancelled while the assets they had shorted rebounded in price. A repeat event, especially at the close, could dramatically erode investor confidence and participation in the capital markets, reducing liquidity and increasing transaction costs.

7.3.1. Theories of the Flash Crash. Initial speculation for the cause of the Flash Crash included so-called fat-finger trading error, a software bug, or a malicious "denial of service" type attack. Yet, no evidence for these explanations has since come to light. The disproportionate impact of the Flash Crash on ETPs led some commentators to draw a connection between the sharp market moves on May 6 and the pricing and trading of these instruments (see, for example, Wurgler 2010). Ramaswamy (2010) examines the operational frameworks of ETFs and relates these to potential systemic risks (for a discussion of the role of leveraged ETFs in the context of end-of-day volatility effects, also see Cheng & Madhavan 2009). In a controversial report, Bradley & Litan (2010) conclude that ETF pricing poses unquantifiable but very real systemic risks. The authors further propose ETF-related reforms and note that, in the absence of such rules, other potentially more severe flash crashes are a virtual certainty.

7.3.2. Flash Crash and fragmentation. The joint report of the Commodities Futures Trading Commission and Securities and Exchange Commission provides a chronology of events on May 6, 2010, and identifies the catalyst for the Flash Crash: Faced with increased volume, the NYSE entered slow trading mode while stocks continued to trade in electronic venues, such as BATS, resulting in price distortions. Liquidity providers began to withdraw their liquidity, given concerns that some trades would be cancelled under the erroneous trade rule. As a result, some market sell orders, including stop loss orders, were executed at pennies.

Although the notion that the Flash Crash arose from an unlikely confluence of factors is reassuring and consistent with the absence of widespread and rapid price declines in any asset class or region in recent decades, individual stocks have also recently experienced "micro" flash crashes. For example, on September 27, 2010, Progress Energy fell almost 90% in price before recovering in the next 5 min for no obvious reason. Unlike May 6, these "micro" flash crashes do not cluster in

time and affect only individual stocks. Nonetheless, they are recent phenomena and suggest the presence of more systematic factors.

It seems likely that prices are more sensitive to liquidity shocks in fragmented markets because imperfect intermarket linkages effectively "thin out" each venue's limit order book. Madhavan (2012) finds strong evidence that securities that experienced greater prior fragmentation were disproportionately affected on May 6, 2010. Per Zhang (2011), high-frequency trading accounts for up to 70% of dollar trading volume in US equities, including ETPs. The increase in high-frequency trading has raised concerns, especially given order-cancellation rates in the region of 90% and the fact that these strategies are not well understood. Using audit-trail transaction data, Kirilenko et al. (2010) examine the behavior of the e-mini S&P 500 stock index futures market during the Flash Crash. They conclude that, while high frequency traders were not the trigger of the Flash Crash, their responses exacerbated market volatility.

This analysis provides insight into why ETPs were differentially affected (ETPs accounted for 70% of equity transactions ultimately cancelled on May 6), even though ETP trading is less fragmented than that of other equities. For ETPs whose components are traded contemporaneously, widespread distortion of the prices of underlying basket securities prices can confound the arbitrage pricing mechanism for ETPs, thus delinking price from value. From a public policy viewpoint, the fact that fragmentation is now at its highest level ever [as shown by Madhavan (2012) using daily tick data from 1994 to 2011] may help explain why the Flash Crash did not occur earlier in response to other liquidity shocks.

8. SYSTEMIC RISK

8.1. Excess Shorting

Several commentators (e.g., Bradley & Litan 2010) have argued that, when ETFs are sold short, aggregate long and synthetic long positions can exceed in total the actual number of outstanding ETF shares. Thus, some argue that an ETF could, in theory, be bankrupt if investors simultaneously redeem their shares, as redemptions would exceed available assets to be redeemed. In reality, on the settlement day, ETF managers release only redemption proceeds against actual delivery of the ETF shares, i.e., delivery versus payment settlement. As a result, "redemption" by parties that do not physically have ETF shares to deliver (because they have lent them to a short seller) will fail. Although the failure of a large number of redemptions could result in a short squeeze, it would more likely result in redemptions failing. The failure of redemptions would not result in real costs to the ETF, instead impacting only accounting entries that are later cancelled.

> **Example 7: Russell Reconstitution** In July 2007, the Russell indexes underwent a large annual rebalancing with massive redemptions from the US iShares Russell 2000 Index ETF (IWM) from APs who chose to handle the tracking risk. Redemptions from IWM essentially equaled the ETF's assets but were then reversed within a few days. Despite the massive redemptions during the rebalance, at the extremes the index moved with only a 1% band of the NAV. Today, the redeeming AP must certify that it has access to ETF shares to deliver upon settlement of the redemption.

8.2. Securities Lending and Counterparty Risk

Another concern often voiced about ETFs regards securities lending and counterparty risk. Securities lending is the practice whereby a security's owner permits a broker to lend it to another

party, typically for shorting. The borrower usually provides collateral to compensate the lender if the borrower fails to return the borrowed security. Securities lending is significant in dollar terms (more than $1.7 trillion), but it is modest at the individual ETF level, where, on average, less than 10% of ETF holdings are on loan. There is presently a 50% aggregate statutory limit on securities lending for ETFs in the United States. For ETFs, securities lending safeguards include the ability to recall loans from borrowers to cover standard and unexpected redemptions and possibly even the liquidation of the borrower's collateral to "buy-in" the nondelivered securities from the market. Securities lending can enhance ETF returns when safeguarded as above, whereas greater loan activity improves liquidity and price efficiency by reducing the costs of expressing negative views through short selling.

8.3. Settlement Failures

The rise in ETF volumes has been accompanied by increased ETF settlement failures at the clearing corporation. Stratmann & Welborn (2012) document a positive relationship between daily ETF settlement failures and short sale volume and cost to borrow ETFs, which they argue is consistent with market makers failing to deliver to avoid paying borrowing costs associated with their short sales. They also argue that ETF settlement failures Granger-cause higher market index volatility as market makers close out fails positions by trade date plus six days (T + 6), concluding that ETFs' failure to deliver could have consequences for market stability

Securities lending is generally done with multiple collateralized borrowers, and ETF managers have the option of handling AP redemptions in whole or partly with cash. In the extremely unlikely event of massive redemptions that exceed the ETF's available securities (not on loan) and loan reallocations are infeasible, the securities are recalled from the borrower. Under standard lending agreements, if the shares are not returned within the standard market settlement cycle, the borrower is in technical default of the lending agreement. Typically for US equity-based iShares ETFs that clear through the Continuous Net Settlement system at the NSCC (National Securities Clearing Corporation), the custodian fronts the cash to the iShares funds as if the settlement had occurred. Any positive rebate rate to the borrower gets reduced to zero. The fund and lending agent continue to share interest earned on the collateral, thus continuing to benefit.

The AP receives cash plus the available deliverable securities from the fund. If the AP has a contractual right to redeem fully in kind for this particular ETF, she may choose to purchase the shares and pass on all related charges to the lending fund (this would generally be done after the settlement date, not within the settlement cycle). Those charges are covered by collateral seized from the borrower. Alternatively, given that the borrower is in technical default of the lending agreement, the lending agent may seize the borrower's collateral to liquidate it and purchase the securities for delivery to the AP.

8.4. Synthetics

Unlike physical ETFs that hold unlevered baskets of securities, synthetic ETFs derive exposure through derivative contracts and face counterparty risk. Synthetics can be appropriate products (for example, if exposures cannot be accessed physically), and if structured well (using multiple counterparties, collateral, etc.), they can minimize counterparty risk. Some synthetic ETFs created by banks in Europe use a single affiliated counterparty, which reduces funding costs but is also a potential conflict of interest. For example, an equity ETF can enter into a total return swap with an affiliated bank (which swaps the total return on the invested portfolio with the return on the underlying index). Normally, the swap counterparty will deliver the index return, but if the

affiliated bank defaults, the investor will face a very different exposure. Variation in the issuer's risk (manifested in credit spreads of the underlying counterparty) will then be reflected in the secondary market price and, hence, the investor's return.

> **Example 8: Manager Discretion** In evaluating discussions around systemic risk, it is very important to understand that ETF managers have many tools to help protect investors: First, ETFs can run tracking error, using the create/redeem baskets to manage liquidity. Second, ETFs can hold a significant share of their portfolio in out-of-index securities. Third, so-called NAV+ creation (where the manager charges a fee to offset transaction cost) is becoming more prevalent. So, creation and redemption is a relatively flexible process with human oversight.

9. CONCLUSIONS

In response to the recent substantial growth of ETFs, investors, regulators, and academics have sought to assess and understand its implications. This article reviews the research to date and develops a model that emphasizes arbitrage as a driver of liquidity and price dynamics. The unified framework here provided is also used to analyze several aspects to the pricing and trading of ETFs including their role in price discovery, the nature of premiums and discounts, performance and tracking relative to benchmark, return autocorrelations, as well as transaction costs and liquidity sourcing in underlying and secondary markets.

The framework is also applied to understand some of the issues related to so-called alternative beta, which is based on model-driven, quantitative portfolio construction techniques, and to active funds that attempt to outperform a benchmark index. Finally, I review several recent policy questions concerning the impact of passive flows on underlying securities, leveraged and inverse ETFs, the role of ETFs in the Flash Crash of May 2010, and issues concerning systemic risk including excess shorting and settlement failures. In sum, ETFs have extended significant benefits to investors and to the functioning of markets that meaningfully outweigh any perceived or actual weaknesses.

DISCLOSURE STATEMENT

The views expressed here are those of the authors alone and not necessarily those of BlackRock, its officers, or directors. This article is intended to stimulate further research and is not a recommendation to trade particular securities or of any investment strategy. Information on iShares ETFs is provided strictly for illustrative purposes and should not be deemed an offer to sell or a solicitation of an offer to buy shares of any funds that are described in this presentation.

ACKNOWLEDGMENTS

I owe a big debt of gratitude to my coauthors Ben Golub, Ira Shapiro, Kristen Walters, Barbara Novick, and Mauricio Ferconi, who helped shape my thinking as reflected herein through our joint partnership on a recent BlackRock Investment Institute "Viewpoints" article. Bob Jarrow, Daniel Morillo, and Antti Petajisto provided many helpful suggestions and improvements. Of course, any errors are entirely my own.

LITERATURE CITED

Arnott RD, Hsu JC, Moore P. 2005. Fundamental indexation. *Financ. Anal. J.* 61(2):83–99

Avellaneda M, Zhang SJ. 2009. *Path-dependence of leveraged ETF returns*. Work. Pap., Courant Inst. Math. Sci., N.Y. Univ.

Ben-David I, Franzoni F, Moussawi R. 2011. *ETFs, arbitrage, and contagion*. Work. Pap. 2011-20, Dice Cent., Ohio State Univ.

Black F. 1972. Capital market equilibrium with restricted borrowing. *J. Bus.* 45:444–55

Bradley H, Litan RE. 2010. *Choking the recovery: Why new growth companies aren't going public and unrecognized risks of future market disruptions*. Work. Pap., Kauffman Found.

Broman MS. 2013. *Excess co-movement and limits-to-arbitrage: evidence from exchange-traded funds*. Work. Pap., Schulich Sch. Bus., York Univ.

Cheng M, Madhavan A. 2009. The dynamics of leveraged and inverse exchange-traded funds. *J. Invest. Manag.* 7(4):43–62

Chow T-m, Hsu J, Kalesnik V, Little B. 2011. A survey of alternative equity index strategies. *Financ. Anal. J.* 67(5):38–57

Clarke RG, de Silva H, Thorley S. 2006. Minimum-variance portfolios in the U.S. equity market. *J. Portf. Manag.* 33:10–24

Da Z, Shive S. 2013. *Exchange-traded funds and equity return correlations*. Work. Pap., Dep. Finance, Univ. Notre Dame

Engle RF, Sarkar D. 2006. Premiums-discounts and exchange-traded funds. *J. Deriv.* 13(4):27–45

Flood C. 2012. ETFs as a driver of US small-cap sector. *Financ. Times*, Feb. 23

Giese G. 2010. *On the performance of leveraged and optimally leveraged investment funds*. Work. Pap., RobescoSAM Indices

Golub B, Novick B, Madhavan A, Shapiro I, Walters K, Ferconi M. 2013. *Viewpoint: exchange traded products: overview, benefits and myths*. Work. Pap., BlackRock Invest. Inst.

Grégoire V. 2013. *Do mutual fund managers adjust NAV for stale prices?* Work. Pap., Finance Div., Univ. B.C.

Haryanto E, Rodier A, Shum PM, Hejazi W. 2013. *Intraday share price volatility and leveraged ETF rebalancing*. Work. Pap., Rotman Sch. Manag., Univ. Toronto

Hasbrouck J. 2003. Intraday price formation in US equity index markets. *J. Finance* 58(6):2375–99

Haugh M. 2011. *A note on constant proportion trading strategies*. Work. Pap., Dep. Ind. Eng. Oper. Res., Columbia Univ.

Hendershott T, Madhavan A. 2014. Click or call? Auction versus search in the over-the-counter market. *J. Finance*. Epub ahead of print; doi: 10.1111/jofi.12164

Jarrow RA. 2010. Understanding the risk of leveraged ETFs. *Finance Res. Lett.* 7:135–39

Kirilenko A, Kyle AS, Samadi M, Tuzun T. 2010. *The Flash Crash: the impact of high-frequency trading on an electronic market*. Work. Pap., Smith Sch. Bus., Univ. Maryland

Lu L, Wang J, Zhang G. 2009. *Long-term performance of leveraged ETFs*. Work. Pap., Guang Hua Sch. Manag., Peking Univ.

Madhavan A. 2012. Exchange-traded funds, market structure, and the Flash Crash. *Financ. Anal. J.* 68(3):20–35

Madhavan A, Morillo D. 2014. The impact of flows into exchange-traded funds: volumes and correlations. *J. Portf. Manag.* In press

Mazza DB. 2012. Do ETFs increase correlation? *J. Index Invest.* 3:45–51

Morillo D, Da Conceicao N, Hamrick J, Stewart S. 2012. Index futures: Do they deliver efficient beta? *J. Index Invest.* 3(2):76–80

Perold AF. 2007. Fundamentally flawed indexing. *Financ. Anal. J.* 63(6):31–37

Petajisto A. 2013. *Inefficiencies in the pricing of exchange-traded funds*. Work. Pap., Stern Sch. Bus., N.Y. Univ.

Ramaswamy S. 2010. *Market structures and systemic risks of exchange-traded funds*. Work. Pap. 343, Bank Int. Settl.

Stratmann T, Welborn JW. 2012. *Exchange-traded funds, fails-to-deliver, and market volatility*. Work. Pap. 12-59, Dep. Econ., George Mason Univ.

Sullivan R, Xiong JX. 2012. How index trading increases market vulnerability. *Financ. Anal. J.* 68(2):70–85
Tuzun T. 2013. *Are leveraged and inverse ETFs the new portfolio insurers?* Work. Pap., Board Gov., Fed. Reserve Syst.
Wimbish W. 2013. Serious health warnings needed for some ETFs. *Financ. Times*, June 23
Wurgler J. 2010. *On the economic consequences of index-linked investing*. NBER Work. Pap. 16376
Zhang F. 2011. *High-frequency trading, stock volatility, and price discovery*. Work. Pap., Sch. Manag., Yale Univ.

Annual Review of
Financial Economics

Volume 6, 2014

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Financial Economics* articles may be found at http://www.annualreviews.org/errata/financial