JOIM
www.joim.com

Journal Of Investment Management, Vol. 14, No. 2, (2016), pp. 86–102

PRICE DYNAMICS AND LIQUIDITY OF EXCHANGE-TRADED FUNDS

Ananth Madhavan and Aleksander Sobczyk[*,†]

Exchange-traded funds (ETFs) have grown substantially in diversity, market significance, and size in recent years. As a consequence, there is increased interest by practitioners in the pricing and trading of these investment vehicles. This paper develops a model to examine ETF price discovery and premium dynamics, and estimates the model individually for 947 US-domiciled ETFs in the period 2005–2014. We find that pricing efficiency varies significantly across funds and is systematically related to cross-sectional measures of liquidity. We provide an illustration of a bond ETF during the financial crisis of 2008 to highlight how apparently dramatic discounts really reflected price discovery when the underlying basket was illiquid in the extreme.

1 Introduction

Exchange-traded funds (ETFs) have grown substantially in size, diversity, and market significance in recent years resulting in increased attention by investors, regulators and academics.[1] The benefits of ETFs—including tax efficiency, transparency, and low cost—are now well understood. Nevertheless, concerns about the pricing and trading of these investment vehicles have also been voiced. Common themes include the transmission of liquidity shocks via arbitrage, excess volatility, and economically significant deviations from the fund's Net Asset Value (NAV).[2] Regulators and policy makers have also recently issued reports examining ETFs from a systemic risk viewpoint. In particular, ETF premiums/discounts are a persistent source of questions regarding "mispricing" because there is no such analogue for an open-ended mutual fund where all transactions occur at NAV at the end of the trading day.

This paper develops a model of ETF price dynamics based on the arbitrage mechanism unique to ETFs to address these questions. Intuitively, both NAV and the market-determined price of the ETF depend on the fund's unobserved intrinsic value. Deviations of price from intrinsic value are corrected over time through arbitrage. By explicitly modeling the temporal dynamics of

*BlackRock, 400 Howard Street, San Francisco, CA 94105, U.S.A.
†The authors are, respectively, Global Head of iShares Research and Director of iShares Research, at BlackRock.

price and NAV we can statistically decompose the ETF's price premium to its NAV into components corresponding to price discovery and transitory liquidity. The model has eight parameters and its general structure admits some interesting special cases. One such case, given concerns about ETF pricing, is where NAV is an accurate representation of intrinsic value while the ETF price reflects noise shocks that may persist for long periods. A polar alternative is one where NAV may exhibit staleness while the ETF price is statistically close to intrinsic value. We estimate the model individually for all 947 US-domiciled ETFs in the period 2005–2014 with a one-year trading history.

The estimates yield several interesting results. First, we derive a natural metric for the speed with which arbitragers act to correct deviations between the ETF's price and intrinsic value. We find that arbitrage speed varies widely across funds and exposures, and is systematically related to cross-sectional measures of liquidity. Specifically, while arbitrage works very quickly for domestic equity funds, the half-life (i.e., the time period over which a given deviation of price from intrinsic value is cut into half) is 6.5 days for international fixed income funds.

The decomposition of the premium into transitory liquidity and price discovery components is important because 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 unobserved intrinsic value. This is often the case during times of market stress. We provide an illustration of a bond ETF during the financial crisis of 2008 to highlight how apparently dramatic discounts really reflected price discovery when the underlying basket was illiquid in the extreme.

Finally, variation in premiums/discounts is closely linked to tracking error and return volatility of the ETF relative to NAV. We show theoretically that ETF return volatility will exceed that of NAV returns when there is staleness in NAV. Some funds, especially comprised of less liquid securities, do exhibit staleness in NAV as manifested by serial correlation in NAV returns. So-called "excess volatility" or tracking error of ETFs can then be explained as a function of pricing conventions.

Our research complements recent analyses of ETFs. In a widely cited study, Ben-David *et al.* (2014) argue that "liquidity shocks in the ETF market are propagated via arbitrage trades to the prices of underlying securities, adding a new layer of non-fundamental volatility." We show that excess volatility may in fact arise due to staleness in NAV and price discovery in the ETF. A related strand of the literature focuses on ETF premiums and discounts. Petajisto (2013) examines the deviation of midquote prices of exchange-traded funds from their Net Asset Values showing they are larger in funds holding international or illiquid securities. Chacko *et al.* (2013) use ETF premiums as a metric for bond market illiquidity. See also Engle and Sarkar (2006) who examine intraday premiums and Tucker and Laipply (2013) who study the time-series properties of select bond ETFs. Our analysis looks across asset classes and exposures, and provides a decomposition of the premium and speed of error correction.

Hasbrouck (2003) uses a vector error correction model to examine the information share of ETFs relative to floor-traded and electronically traded futures (E-minis) contracts for major US equity indexes. He finds that for the S&P 500 and Nasdaq 100 indexes, most price discovery occurs in the E-mini contracts. Our focus, however, is not on the relative contributions of different exposure vehicles, but a fund-specific analysis of the speed of price discovery, broken down by asset classes and exposures.

Finally, there is a growing literature on systemic risk (see Ramaswamy, 2010; International Monetary Fund, 2014; Office of Financial

Research, 2013) with a focus on ETFs and their functioning in stressed markets when liquidity is scarce. Our results highlight the positive role of ETFs in price discovery in such times. Golub *et al.* (2013) provide a comprehensive overview of ETFs and address many of the misconceptions around ETFs as a source of systemic risk.

This paper is organized as follows: Section 2 provides the required institutional background; Section 3 develops a model that we use to analyze ETF premiums and price discovery; Section 4 estimates the model individually for all US-domiciled ETFs from 2005 to 2014, focusing on the speed of arbitrage across asset classes and exposures; Section 5 applies the model to the estimation of true premiums and also the behavior of bond ETFs in the Financial Crisis; and finally, Section 6 concludes and offers some directions for future research.

2 Overview of ETF pricing and trading

2.1 Creation/redemption mechanism

The common characteristic of ETFs is that they are traded intraday on an organized exchange.[3] Although simple, this description ignores many important differences between ETFs and other pooled investment vehicles.[4] In particular, ETFs share elements of both open- and closed-end mutual funds. For an open-ended mutual fund, transactions occur only at the end of the day and only at NAV. In contrast to open-ended mutual funds but like closed-end funds, ETFs trade during the day in the secondary market at prices that can deviate from NAV. Further, unlike open-ended mutual funds, ETFs issue and redeem shares only in a minimum size (creation unit) and only with market making firms known as Authorized Participants (APs). The creation/redemption mechanism in the ETF structure allows the number of shares outstanding in an ETF to expand or contract based on demand from investors. The creation–redemption mechanism means that liquidity can be accessed 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, a factor especially important for large investors.

Transparency is the key to pricing. 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. The 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.[5] Like other 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. Although ETF shares are created or redeemed at the end of each trading day, this is an accounting issue because the AP will typically lock in its profits intraday by selling the higher-priced asset while simultaneously buying the lower-priced asset. For example, when an ETF is trading at a premium to an AP's estimate of value, the AP may choose to deliver the creation basket of securities in exchange for ETF shares, which in turn it could elect to sell or keep. The creation–redemption mechanism works through arbitrage to keep the ETF's price close to the intrinsic value of an ETF's holdings in the underlying market.

2.2 Trading and pricing

Deviations of price from NAV do not necessarily imply the existence of arbitrage opportunities. The ETF provider contracts with market data vendors (or other third-parties) to calculate and publish NAV based on past prices. Market data

vendors can adjust last prices/quotes for new market information.[6] For example, they may make adjustments to prevent "market timing" in international and fixed income funds where previous prices or quotes may be recorded with a substantial delay. Approaches to fair valuation vary and there is no accepted standard to adjusting stale prices. Indeed, Grégoire (2013) finds evidence that mutual funds do not fully adjust their valuations to reflect fair value and returns remain predictable.

Vendors also provide an Intraday Indicative Value (IIV) that is disseminated at regular intervals during the trading day, typically every 15 seconds.[7] This value is usually based on the most recent (possibly stale) trade, and not on the midpoint of the current quote on each portfolio component. So, if the fund holds Japanese stocks, say, the closing price (or quote) from Tokyo is used throughout the US trading day and a foreign exchange adjustment is made for any change in the yen/dollar relationship since the Tokyo markets are closed. For fixed income funds, IIV can often be quite misleading as vendors usually update their NAV towards the close. For this reason, IIV is not useful for APs and arbitragers for trading purposes. Rather, these market participants will use their own proprietary models and data to estimate the underlying value of the ETF.

Large institutions, market makers, and proprietary traders will incorporate information from a variety of sources in their internal evaluations. In the case of an international fund holding Japanese stocks, for example, this may include Japanese ADRs and the Nikkei futures in the US. For bond funds, information on key interest rates and spread changes are relevant, as well as transaction data from sources such as TRACE. Market makers can use the current bids and offers for portfolio securities, rather than the last sales in their calculations of an ETF's current portfolio value. More importantly, they have proprietary knowledge of the size and side of much of the current order flow in the ETF. Proprietary traders (including the Authorized Participants who actually execute creation/redemption transactions) often work with market makers in their joint risk management. In conclusion, the price and NAV of an ETF are distinct values, and both may differ from the expected intrinsic value of the portfolio for the reasons discussed above. In what follows, we lay out a simple framework that captures these elements.

3 Analytical framework

3.1 Price, NAV and expected value

From the above discussion, to understand ETF price and premium dynamics we need to model three distinct values: The ETF secondary market price, NAV, and the (unobserved) expected value of the ETF. We can interpret the price/value variables as represented in natural logs, so that first differences are continuously compounded returns. The expected value of the asset at a point in time t is denoted by v_t. As discussed above, we do not observe expected value. Changes in conditional expectations are innovations, so we model v_t as a random walk with drift:

$$v_t = v_{t-1} + r_t, \tag{1}$$

where $r_t \sim (\mu_r, \sigma_r^2)$ is a stochastic return with mean μ_r and variance σ_r^2.

The ETF price is expected value, v_t, plus a shock u_t that can take on positive or negative values:

$$p_t = v_t + u_t. \tag{2}$$

We refer u_t as the "true premium" that reflects transitory price pressure. Note that $|u_t|$ may be larger or smaller than the quoted half bid–ask spread. Large trades may trade outside the quotes while others can negotiate inside the quotes.

We model the true premium (see also Poterba and Summers, 1988) as:

$$u_t = \psi u_{t-1} + \varepsilon_t, \tag{3}$$

where $\varepsilon_t \sim (\mu_\varepsilon, \sigma_\varepsilon^2)$ is a liquidity shock and ψ is the autocorrelation coefficient. This representation is intuitive—true pricing errors are serially correlated but are corrected over time. Lower values of ψ imply faster error correction, with the extreme case of $\psi = 0$ implying that errors are corrected immediately. Equation (3) can be motivated by a model where investors respond similarly—but not simultaneously—to news and macroeconomic shocks. The impact of flows on prices is offset by arbitragers who trade in the opposite direction to capture the true premium. Inventory and redemption costs, risk aversion, and price impact in both the ETF and underlying markets explain why pricing errors are not instantly eliminated by arbitrage. So, ψ is positively related to dealer inventory costs and uncertainty over fundamentals, and is positively related to the autocorrelation in exogenous flows.

3.2 Premiums and discounts

The observed premium (or discount) is defined as the deviation of the ETF price from the Net Asset Value (NAV) of the fund:

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

where n_t is the NAV at time t. (In our subsequent empirical work, we will use the published end-of-day NAV for n_t.) We model NAV as a weighted average of current value (possibly with pricing noise) and past NAV:

$$n_t = (1 - \varphi)v_t + \varphi n_{t-1} + w_t. \tag{5}$$

Here $0 \leq \varphi \leq 1$ captures possible staleness and $w_t \sim (\mu_w, \sigma_w^2)$ is an error term that reflects microstructure effects (such as the bid pricing convention for fixed income ETFs). This formulation captures several interesting cases.[8] In particular, if $\varphi = 0$ and $\mu_w \sim 0$, Equation (5) implies that on average NAV equals expected value and there is no staleness. Consequently, from Equation (2), premiums and discounts in the ETF reflect transitory liquidity shocks. Another case of interest is if $\varphi > 0$, so NAV exhibits staleness. For example, illiquid fixed income assets are often traded in dealer markets, where current quotations may not be available on a timely basis or may not be representative of current market conditions. Pricing providers may use proprietary processes such as matrix pricing to adjust stale bond quotes, but it is not clear if these adjustments can capture all the relevant changes in spreads or rates for a given bond. Alternatively, $\varphi < 0$ would correspond to over-reaction where the pricing provider places too much weight on new information.

Any pricing noise will be reflected in w_t. For example, for bond ETFs, NAV is based (by regulation) on a bid pricing convention so that w_t has mean (denoted by μ_w) equal to negative half of the average bid–ask spread of the underlying bonds. Dispersion in w_t thus reflects variation in the bid–ask spread due to changes in volatility or liquidity. It also reflects the point above that different investors may execute at different prices.

The profit of a market maker is not the ETF's premium but the deviation of price from expected value, u_t. While creations and redemptions are at 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. Using the definition of the premium $\pi_t = (p_t - n_t)$ we get:

$$\pi_t = (v_t + u_t) - n_t. \tag{6}$$

We can solve[9] Equation (6) to express the premium as:

$$\begin{aligned}\pi_t = \varphi(r_t + \varphi r_{t-1} + \cdots) \\ + (1-\varphi)(w_t + \varphi w_{t-1} + \cdots) + \varepsilon_t \\ + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \cdots\end{aligned} \tag{7}$$

This expression shows the composition of the ETF's premium into three terms: (a) *Price Discovery*, the product of the staleness factor and a weighted average of past fundamental returns; (b) a weighted average of past NAV *pricing noise*; and (c) *Transitory Liquidity*, captured by a weighted average of past liquidity innovations. When NAV is current and pricing noise is minimal, $\varphi = 0$ and the premium is $\pi_t = \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \cdots$ or a weighted average of all past flow shocks. Higher values of ψ, which correspond to greater autocorrelation in flows and less efficient arbitrage, imply that past shocks have a greater effect on premiums.

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 NAV implies a positive mean for the premium. From an investor perspective, a deviation of price from NAV need not imply that the ETF is mispriced, an important point as some investors may avoid buying at a premium or selling at a discount.

3.3 Volatility and shock propagation

Some commentators have emphasized cases where ETF return volatility is greater than that of NAV returns. However, it is straightforward to explain this with the model. Ignoring for simplicity of any microstructure noise w_t, we can express NAV returns as:

$$n_t - n_{t-1} = (1-\varphi)r_t + \varphi(n_{t-1} - n_{t-2}). \tag{8}$$

The variance of NAV returns is:

$$\sigma^2(n_t - n_{t-1}) = \sigma_r^2 \frac{(1-\varphi)^2}{1-\varphi^2}. \tag{9}$$

So, NAV return variance is a fraction of fundamental variance σ_r^2. This fraction $\frac{(1-\varphi)^2}{1-\varphi^2} < 1$ when $\varphi > 0$ and decreases with staleness. If there is no staleness, $\varphi = 0$ and $\sigma^2(n_t - n_{t-1}) = \sigma_r^2$. Note that microstructure noise can imply that NAV returns have greater volatility than fundamental returns, but we do not expect this effect to be economically significant.

The 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[10] to be:

$$\sigma^2(p_t - p_{t-1}) = \sigma_r^2 + 2\sigma_\varepsilon^2(1+\psi)^{-1}. \tag{10}$$

The first term is fundamental return variance and the second reflects variance from liquidity shocks. So, if there is staleness in NAV, ETF return volatility will exceed that of NAV returns. The differences are greater, the higher the degree of staleness 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 $\sigma^2(r_{0,T}) = \sigma_r^2 T$. The variance of the microstructure shock difference component of ETF returns, $\sigma^2(u_t - u_{t-1}) = 2\sigma_\varepsilon^2(1+\psi)^{-1}$ and does not scale with time, so tracking difference narrows over longer intervals.

A related issue concerns whether ETF trading propagates volatility into the underlying securities, increasing transitory volatility and having other detrimental effects.[11] Ben-David *et al.* (2014) argue that ETF liquidity shocks are propagated via arbitrage trades to underlying markets,

creating non-fundamental volatility. Their metric of ETF "mispricing" is the volatility of the daily premium or $\sigma(\pi_t)$ in our notation. Using our model, the premium can be written as (we ignore the impact of microstructure noise in NAV for simplicity):

$$\pi_t = \frac{\varphi}{1-\varphi}(n_t - n_{t-1}) + u_t, \qquad (11)$$

where the error term u_t follows a first-order autoregressive process. This expression makes it clear that "mispricing" is contemporaneously correlated with return volatility, so a regression of broader market volatility on "mispricing" will necessarily show statistical significance. In other words, an alternative explanation to shock propagation is that after an innovation in fundamentals, ETFs lead price-discovery, NAV "catches up" over time, with no evidence of causality.

3.4 Speed of arbitrage

We can compute speed of mean reversion in terms of the half-life, i.e., the time horizon need on average to halve an error. Since the "true" premium is a stationary process, the expected premium h periods ahead from period t is $E[u_{t+h}] = \psi^h u_t$. So, if the half-life is h, then $E[u_{t+h}] = 0.5u_t$ and the half-life is thus:

$$h = \left(\frac{\ln(0.5)}{\ln(|\psi|)}\right). \qquad (12)$$

Intuitively, arbitragers act to correct pricing errors so their trading causes a convergence of price to expected value. Lower values of ψ imply fast arbitrage and less serial dependence in pricing errors. As market makers are risk averse and face market impact costs, price does not instantly revert back to expected value (net transaction costs), but rather corrects over time. This is also the case if dealer capital is limited. Note that ψ also reflects the correlation in flow from period to period. If flows are strongly temporally correlated and investors crowd on the same side of the market, the ETF will trade at prices above or below fundamental value for multiple periods.

4 Empirical estimation

4.1 State-space representation

Empirical estimates of the model at the individual fund level are of economic interest because they provide insights into the efficiency of the arbitrage mechanism measured by the estimated "true" deviations between price and (unobserved) value through ψ. We are interested, for example, in learning why some funds exhibit greater price efficiency than others, and the length of time it takes for pricing errors to be corrected. The degree to which a fund's NAV is stale (as captured by a positive value φ) is also interesting as staleness implies predictability in NAV returns, and as shown here, is related to the difference in volatility in ETF price-based returns versus NAV returns. Further, the model estimates let us decompose the premium to estimate the portion of the average premium attributable to liquidity versus price discovery. While market makers and others in the ecosystem with proprietary models, data, and views on flow may have accurate views of the true premium and trade on this, many other investors do not have such data and may sometimes postpone a purchase or sale based on observed premiums or discounts. As noted earlier, the existence of a premium or discount need not imply any mispricing, as we illustrate below.

There are several possible approaches one can take to estimate the model given time-series data on prices and NAV for a given fund. As shown above, price and NAV returns depend on the parameters, so armed with expressions for return volatility and autocorrelations, we have enough moment conditions to estimate the model's parameters. Alternatively, one can estimate the model using a multivariate state-space

(Kalman filter) representation. We adopt the state-space representation because it aligns with the model directly and lets us explicitly estimate the unobserved true premium, u_t. The Kalman filter is the best possible (optimal) estimator for a large class of problems where we want to make inference based on observations of noisy signals, and is used in a variety of real-world applications where estimates are based on mechanical, optical, acoustic, or magnetic sensor data. See, e.g., Hamilton (1994) for further details.

The state-space representation consists of two elements: (1) The measurement (or observation) equation with price and NAV, expressed as a function of the unobserved state vector (expected value); and (2) The transition (or state) equation which expresses the dynamics of the state vector. The Kalman filter works by using the measurement equation to create dynamic forecasts of the state vector, much like a conventional regression. In our case, the measurement equation is:

$$\begin{bmatrix} p_t \\ n_t \end{bmatrix} = \begin{bmatrix} \psi p_{t-1} \\ \varphi n_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & -\psi \\ 1-\varphi & 0 \end{bmatrix} \begin{bmatrix} v_t \\ v_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ w_t \end{bmatrix}. \quad (13)$$

The transition equation describes the random-walk process for unobserved expected value:

$$\begin{bmatrix} v_t \\ v_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_{t-1} \\ v_{t-2} \end{bmatrix} + \begin{bmatrix} r_t \\ 0 \end{bmatrix}. \quad (14)$$

There are 8 parameters to be estimated for each fund: The coefficients φ and ψ (staleness and efficiency) and the means and variances of the shocks ε_t, w_t, and r_t. As an aside, the framework in Equations (13)–(14) can be extended to explicitly include ETF flows, data that are publicly available. Specifically, we can model flows as serially correlated and related to past return innovations. We expect flows to be positively related to the liquidity shock ε_t reflecting market impact. Such an augmented model has 3 additional parameters of economic interest: flow autocorrelation, return sensitivity, and the market impact coefficient.

4.2 Descriptive statistics

We estimate the models (13)–(14) using maximum likelihood for the universe of all US-domiciled ETFs from January 1, 2005 to January 31, 2014. We restrict attention to physically-backed ETFs on equities and fixed income, excluding exchange-traded notes (which are really debt instruments), leveraged and inverse products, and other synthetic funds. Estimation uses daily closing prices and NAVs sourced from BlackRock and Bloomberg, log-transformed to be consistent with the model. We require that funds have at least 252 consecutive trading days of history. Note that we do not limit the sample to funds that are listed at the end of the sample period, but simply require a year's continuous trading.

Summary statistics for the sample in terms of assets under management and trading characteristics are provided in Table 1, broken down by asset class and exposure. For the sample, total assets under management (AUM) represent almost $1.5 trillion in 947 funds, which is comprehensive in the sense that total AUM in US ETPs was $1.86 trillion as of June 2014.[12] The great majority of the funds and assets are in domestic equity ETFs, followed by international equity. Domestic equity funds are the most liquid by conventional measures such as the spread, trading frequency, or dollar value traded. Consistent with Petajisto (2013), international equity funds have the largest absolute premium of 73.4 basis points, versus 48.7 for all funds. Note that there is considerable variation in absolute premiums over time at the individual fund level.

The model has 8 parameters implying a total of 7,576 estimates, so we report summary measures

Table 1 Descriptive statistics.

	Equity		Fixed income		
	Domestic	International	Domestic	International	All funds
Number of funds	387	403	113	44	947
Total AUM ($MM)	892,804	382,021	191,240	21,998	1,488,063
Average AUM ($MM)	2,307	948	1,692	500	1,571
Average number sample days	1,581	1,224	1,116	931	1,343
Average ADV ($MM)	106.0	26.7	28.6	7.1	57
Average trades per day	2,854	1,635	1,029	350	1,980
Average bid/ask spread (bps)	16.6	52.4	28.6	54.8	35.4
Average premium (bps)	−1.8	18.3	18.3	16.7	10.0
Average absolute premium (bps)	23.8	73.4	37.6	69.8	48.7

Source: Bloomberg and BlackRock, 1/1/2005–1/31/2014. The figures for assets under management (AUM) are taken from the last trading day of the sample. For bid–ask spreads, average daily volumes (ADV) and number of trades, we use the past year for computation, and report the unweighted means by asset class and exposure.

of the individual estimates. We focus on the two parameters of greatest interest namely the staleness coefficient and the speed of arbitrage. Also estimated, but not reported here, are the standard deviations of the three innovation terms (NAV, price, and value) and their respective means. Table 2 reports the mean (both simple and AUM-weighted) and median estimates of the staleness and arbitrage speed parameters and the standard deviation of these estimates, broken down by asset class and exposure category. We also report the fraction of the estimated coefficients that are significantly greater than zero.

Observe that the staleness parameter estimates increase as we move from the most liquid asset classes (domestic equity) to the less liquid asset classes (fixed income), consistent with our prior observations. For domestic equity, staleness is, in general, both economically and statistically insignificant, in line with intuition. The speed of arbitrage (which is measured *inversely* by ψ) increases with liquidity, ranging from a median of 0.20 in domestic equity to 0.90 in international fixed income. The corresponding half-life using Equation (12) for correcting a given unobserved pricing error is 0.43 to 6.56 days, respectively. Again, these estimates are consistent with our intuition that domestic equity exhibits relatively low staleness compared to international fixed income.

4.3 Multivariate analysis

A multivariate analysis allows us to examine how arbitrage speed, staleness, and pricing errors vary across funds while jointly accounting for fund size, trading activity, and exposure characteristics. We regress fund-level estimates of arbitrage speed, staleness, and the standard deviation of noise shocks on log AUM, log of dollar ADV, and indicator variables taking the value 1 if the fund has a fixed income or international focus (i.e., geography is non-US), respectively, and 0 otherwise. The coefficient estimates with associated t-statistics in parentheses are summarized in Table 3.

Fixed income and international funds exhibit significant NAV staleness and slower error correction as shown by positive coefficients in Models 1 and 2, respectively. Fund size is associated with

Table 2 State-space model estimates.

		Equity		Fixed income		
		Domestic	International	Domestic	International	All funds
Number of funds		387	403	113	44	947
NAV staleness coefficient (φ)	Mean	−0.08	0.15	0.40	0.41	0.10
	Median	−0.05	0.15	0.45	0.33	0.00
	Std. dev.	0.11	0.18	0.35	0.26	0.26
	Wtd. mean	−0.02	0.22	0.51	0.32	0.12
	Significant > 0	0.03	0.74	0.83	0.95	0.47
Arbitrage speed parameter (ψ)	Mean	0.24	0.43	0.61	0.79	0.39
	Median	0.20	0.44	0.71	0.90	0.34
	Std. dev.	0.23	0.50	0.33	0.21	0.41
	Wtd. mean	0.28	0.19	0.69	0.68	0.32
	Significant > 0	0.80	0.77	0.96	1.00	0.82

Source: Bloomberg and BlackRock data, 1/1/2005–1/31/2014. Weighted means are based on AUM weights. Significance is the fraction of the estimate that is greater than zero, based on a one-tailed t-test at the 5% level.

Table 3 Multivariate regression models.

Variables	(1) NAV staleness coefficient (φ)	(2) Arbitrage speed parameter (ψ)	(3) Std. Dev. liquidity shocks ($\sigma_\varepsilon \times 100$)
Intercept	−0.22	0.18	0.79
	(−5.42)	(2.23)	(13.68)
Log AUM	0.03	0.01	−0.09
	(3.75)	(0.90)	(−8.12)
Log Dollar ADV	−0.01	−0.02	0.03
	(−1.50)	(−1.44)	(2.92)
Fixed Income	0.40	0.36	−0.14
	(23.37)	(10.52)	(−5.67)
International	0.23	0.19	0.07
	(16.85)	(6.83)	(3.90)
Adjusted R-square	0.44	0.13	0.28
F-value	180.01	35.61	88.24
Degrees of freedom	918	918	918

Source: Bloomberg and BlackRock data, 1/1/2005–1/31/2014. t-Statistics are in parentheses. The figures for assets under management (AUM) are taken from the last trading day of the sample; for dollar average daily volume, we use the past year as the period for computation.

greater staleness, perhaps because larger funds track broad indexes with many constituents and are more difficult to mark to market. Finally, the volatility of liquidity shocks in Model 3 is significantly lower for larger funds and fixed income funds, but is higher for international funds.[13]

5 Analysis of premiums, arbitrage, and price efficiency

5.1 True premiums

A case application of the model to a particular fund yields valuable insights. We selected the iShares iBoxx High-Yield Corporate Bond ETF (HYG) during the 2008–2009 Financial Crisis to illustrate how the model can be used to understand reported premiums or discounts. Using the estimated model, we can recover the implicit state vector (i.e., the time series of $\hat{v}_t$) which is plotted in Figure 1 along with the daily closing price and NAV.[14] Both price and NAV move closely together until the start of the crisis, and the estimated state variable lies between these values. In September 2008, price moves down sharply relative to NAV, increasing the discount, and the staleness in NAV is apparent. In March 2009, when the market recovers, price leads NAV upward. In this period, the recovered state variable tracks price quite closely, similar to the results reported by Tucker and Laipply (2013).

Consistent with these observations, a regression of HYG's daily NAV returns in this period on the previous day's return yields a coefficient of 0.551 (t-value of 10.82) while the corresponding coefficient for ETF daily returns was slightly negative and statistically insignificant.[15] Recall further that the model shows that ETF returns will exhibit greater volatility than NAV returns if there is staleness, and indeed, the volatility of HYG daily returns from April 2007 to end-December 2013 is approximately twice that of NAV returns, 93 versus 41 basis points.

The estimated residuals $\{\hat{u}_t\}$ yield a time series estimate of the "true" premium for a given fund at any point in time. We plot in Figure 2 the estimated true premium $\{\hat{u}_t\}$ against the observed

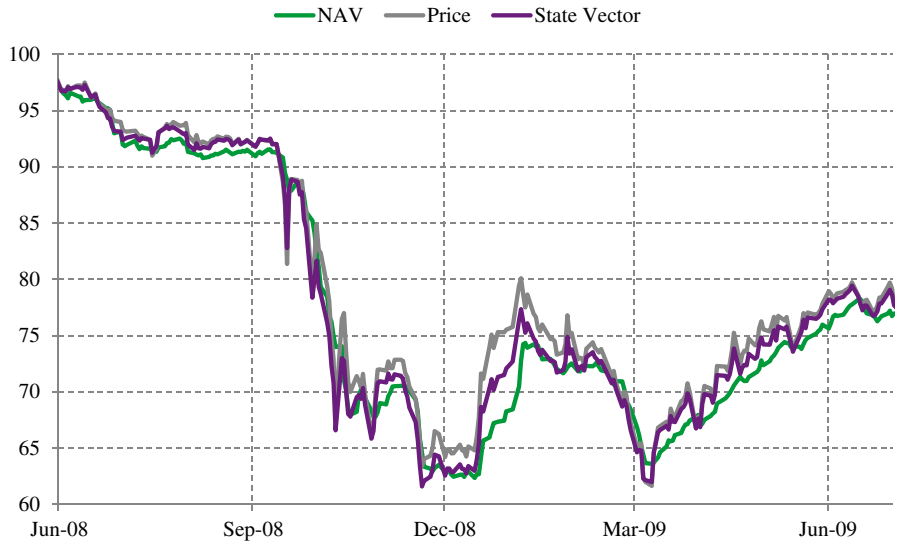

Figure 1 iShares High-Yield Corporate Bond ETF (HYG) prices, NAV, and state vector from June 2008 to June 2009.

Source: Bloomberg (price and NAV) and BlackRock (State Vector Estimate), 6/1/2008–6/30/2009.

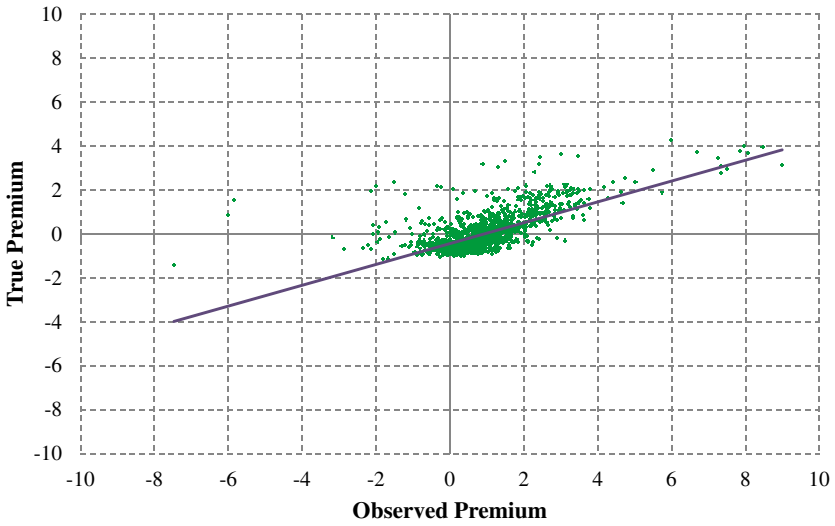

Figure 2 Observed vs. true premiums for iShares High-Yield Corporate Bond ETF (HYG).

Source: Bloomberg and BlackRock data, 4/13/2007–12/31/2013.

premium $\{\pi_t\}$ for HYG in the period April 13, 2007 to December 31, 2013, where consistent with the model, both variables are log-transformed. A regression of the true premium on the observed premium yields a slope coefficient of 0.47 (t-value 38 and adjusted R-square of 0.461), indicating that roughly half the observed premium on average is due to price discovery.

Why is this relevant? Many investors in ETFs including Financial Advisors, retail clients, hedge funds, etc. are concerned that buying (selling) at large premium (discount) is unfavorable. While market makers typically have the tools to form a good estimate of intrinsic value, their data/models are proprietary. The Kalman filter approach can be used to produce a step-ahead estimate of the true premium and the speed with which it is corrected for a given fund on a daily or intraday basis.

5.2 Price discovery

Recall that the premium at any point in time can be expressed as $\pi_t = p_t - n_t = (p_t - v_t) + (v_t - n_t) = u_t + (v_t - n_t)$. In the extreme case where the ETF price always reflects true value, the first term u_t will be zero and the entire premium reflects staleness in NAV. Alternatively, if there is no NAV staleness, then $n_t = v_t$ and the premium entirely reflects the shock u_t to the ETF price through the secondary market. The standard error of u_t is $\sigma_u = \sigma_\varepsilon/\sqrt{(1-\psi^2)}$. This term is positively related to the volatility of transitory liquidity shocks σ_ε and is also increasing in ψ, meaning that more efficient arbitrage means smaller residuals. We define the price discovery component of the premium as the portion of total variance that is not attributable to transitory noise shocks, that is:

$$D = 1 - (\sigma_{\mathrm{u}}/\sigma_\pi)^2, \tag{15}$$

where σ_π is the standard deviation of the observed premium.

The price discovery component D is shown in Table 4 by asset class and AUM quintiles with 1 being the largest and 5 the smallest. It is clear that the price discovery component declines as fund size drops in all four categories of asset class and exposure. In other words, transitory liquidity shocks constitute a larger fraction of the premium for smaller, less actively traded funds. The estimates make intuitive sense in that roughly 74% of

the variation in premiums for large international funds is due to price discovery. That figure is lowest for small international fixed income funds. From a practical perspective, the results confirm our intuition that for the most active funds, where AUM is the largest, apparent premiums or discounts reflect staleness (or pricing errors) in NAV as opposed to transitory liquidity pressures in the ETF market.

5.3 ETFs in the financial crisis

Anecdotal evidence suggests that ETFs played an especially important role during the financial crisis when liquidity in underlying securities, particularly in fixed income, was scarce. To examine this issue, we estimated the model during the financial crisis (from 1/1/2008 to 12/31/2009) and post-financial crisis. We constrained the analysis to the common set of funds for both periods with a full year of trading history (252 trading days), leaving us with 484 funds.

Observe that the NAV staleness coefficients in Table 5 are higher in the crisis period for fixed income funds, perhaps reflecting the decline in trading activity, particularly in individual corporate bonds. Interestingly, the arbitrage speed

Table 4 Estimation of price discovery component.

Asset class	Exposure	AUM quintile	Number of funds	Total AUM ($MM)	Price discovery component (D)
Equity	Domestic	1	78	801,146	0.59
		2	78	64,815	0.49
		3	77	19,804	0.41
		4	77	5,940	0.32
		5	77	1,099	0.29
	International	1	81	347,338	0.74
		2	81	26,107	0.64
		3	81	6,541	0.53
		4	80	1,674	0.43
		5	80	360	0.25
Fixed income	Domestic	1	23	162,123	0.55
		2	23	18,958	0.40
		3	23	7,230	0.50
		4	22	2,409	0.46
		5	22	520	0.45
	International	1	9	18,077	0.31
		2	9	2,556	0.33
		3	9	987	0.50
		4	9	306	0.22
		5	8	73	0.17
All funds			947	1, 488, 063	0.46

Source: Bloomberg and BlackRock data, 1/1/2005–1/31/2014.

Table 5 State-space model estimates during and after financial crisis.

		Equity		Fixed income		
		Domestic	International	Domestic	International	All funds
Number of funds		253	186	34	11	484
A. Crisis period (January 2008–December 2009)						
NAV staleness	Mean	−0.08	0.17	0.35	0.51	0.06
coefficient (φ)	Median	−0.04	0.21	0.30	0.30	−0.02
	Std. dev.	0.08	0.17	0.29	0.32	0.22
	Wtd. mean	−0.03	0.22	0.41	0.51	0.09
	Fr. significant > 0	0.02	0.73	0.88	1.00	0.37
Arbitrage speed	Mean	0.22	0.23	0.57	0.74	0.26
parameter (ψ)	Median	0.21	0.23	0.65	0.78	0.25
	Std. dev.	0.25	0.46	0.28	0.23	0.36
	Wtd. mean	0.26	0.06	0.68	0.80	0.25
	Fr. significant > 0	0.66	0.60	1.00	1.00	0.67
B. Post-crisis period (January 2010–January 2014)						
NAV staleness	Mean	−0.04	0.19	0.29	0.43	0.09
coefficient (φ)	Median	−0.02	0.19	0.12	0.38	0.00
	Std. Dev.	0.06	0.14	0.32	0.25	0.19
	Wtd. Mean	−0.01	0.23	0.29	0.32	0.09
	Fr. Significant > 0	0.04	0.74	0.84	0.95	0.48
Arbitrage speed	Mean	0.27	0.29	0.58	0.86	0.31
parameter (ψ)	Median	0.26	0.41	0.71	0.90	0.30
	Std. Dev.	0.20	0.70	0.35	0.13	0.48
	Wtd. mean	0.36	−0.04	0.54	0.85	0.29
	Fr. Significant > 0	0.83	0.77	0.96	1.00	0.83

Source: Bloomberg and BlackRock data, 1/1/2008–1/31/2014.

parameters are consistently smaller during financial crisis (Panel A) than in the post-crisis period (Panel B), supporting the view that ETFs serve as price discovery vehicles during times of market stress.

6 Conclusions

Premiums and discounts in ETFs are a persistent source of questions from policy makers and practitioners largely because there is no such analogue for an open-ended mutual fund. Specifically, why do ETFs trade at premiums or discounts, and how does this affect tracking error? Why are ETF returns sometimes much more volatile than the volatility of returns of their underlying indexes? How does ETF pricing and liquidity work in times of market stress? Do large premiums/discounts at these times reflect pricing errors? Are these persistent and if so, is this an arbitrage opportunity? How should investors incorporate deviations from net asset value into their investment decisions? Should they avoid buying funds trading at steep premiums or selling

funds at a significant discount? And for policy makers, do ETFs on less liquid asset classes such as high-yield bonds pose systemic risks? Is there any evidence from the crisis on how these investment vehicles performed?

This paper develops and tests an 8-parameter model of ETF price dynamics to better understand these questions. We estimate the model individually for each of 947 US-domiciled equity and fixed income ETFs from 2005 to 2014. Our major findings are as follows:

- The observed fund premium or discounts can be decomposed into price discovery and transitory liquidity components. These components vary systematically across asset class, exposure, and fund size.
- The Net Asset Value of international equity funds and bond ETFs, particularly smaller funds, can exhibit staleness, but this is largely insignificant for domestic equity funds. Staleness arises because NAV does not fully capture current valuations, and the lag in adjustment can give rise to economically significant premiums or discounts, especially in times of market stress. ETF return volatility will be greater than that of the underlying NAV returns if NAV is stale.
- We estimate the speed with which unobserved pricing errors in a given fund are corrected through the arbitrage mechanism. Arbitrage acts quickly to correct pricing errors for domestic equity funds, with a half-life of 0.43 days versus 6.56 days for international fixed income funds.
- Apparently large discounts to Net Asset Value in periods of bond market stress such as the financial crisis reflects efficient pricing, not illiquidity. This result should mitigate concerns that ETFs are the source of additional volatility or of systemic risk.
- The cross-sectional findings provide strong evidence that observed premiums largely reflect price discovery, particularly for ETFs with constituents trading outside of US market trading hours.

The framework established here can be extended in several directions. First, we can include ETF flows by explicitly modeling their dependence on past flows and return innovations and linking flow innovations to transitory price pressure. Second, we can compare the "top down" state-space estimates with a "bottoms up" security-by-security valuation to gain a deeper understanding of the limits of arbitrage. Third, we can make the model more realistic by incorporating cross-correlations into the shocks terms and time-dependence into the parameters. These, however, are subjects for future research.

In conclusion, ETF pricing dynamics are driven by arbitrage, and a deeper understanding of this key mechanism can help practitioners better utilize these powerful tools for gaining a wide range of diversified exposures at low cost.

Notes

[1] As of March 2015, global ETFs represented over $3 trillion in total net assets (Source: BlackRock). Sullivan and Xiong (2012) note that while passively managed funds represent only about one-third of all fund assets, their average annual growth rate since the early 1990s is 26%, double that of actively managed assets.

[2] Dieterich and Cui (2014) state "Large swings in U.S. government-bond prices are renewing investor scrutiny of whether exchange-traded funds, or ETFs, are boosting market volatility." Wimbish (2013) cites concerns that ETFs may "cause additional market-wide systemic problems because of the arbitrage opportunities they produce."

[3] The ETF sponsor originates the fund and selects its investment objective. The great majority of ETFs are index-based, where the sponsor chooses both an index and a method of tracking its target index. See Gastineau (2002) for a detailed description.

[4] Exchange-traded funds (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.

[5] In the case of cash redemptions, transaction charges resulting from investing or raising the cash are absorbed by the AP and not the ETF, unlike open-ended mutual funds. 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.

[6] There is a further subtlety for fixed income funds where industry convention is to value using the bid price, a point we return to later.

[7] This estimate also is also referred to as an Indicative Net Asset Value (INAV) or Intraday Optimized Portfolio Value (IOPV).

[8] An equivalent and intuitive representation of NAV in Equation (5) is as exponentially weighted average of noisy estimates of current and past values where the weight on the j-lagged price is $(1-\varphi)\varphi^j$ so that $n_t = (1-\varphi)\sum_{j=0}^{\infty}\varphi^j(v_{t-j}+\breve{w}_{t-j})$ where $\breve{w}_t = w_t/(1-\varphi)$ is the rescaled noise term.

[9] From the autoregressive formulation of liquidity shocks we have $u_t = \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \cdots$ so the premium is $\pi_t = [1-(1-\varphi)(1-\varphi L)^{-1}](v_t + w_t) + u_t$. As the return $r_t = (1-L)v_t = v_t - v_{t-1}$ the expression follows.

[10] Let L denote the lag operator, i.e., $Lu_t = u_{t-1}$. Then, $u_t = (1-\psi L)^{-1}\varepsilon_t = \varepsilon_t + \psi\varepsilon_{t-1} + \psi^2\varepsilon_{t-1} + \cdots$ The shock change $(u_t - u_{t-1}) = \varepsilon_t + (\psi - 1)(1-\psi L)^{-1}L\varepsilon_t$, so $\sigma^2(u_t - u_{t-1}) = \sigma_\varepsilon^2(1 + \frac{(1-\psi)^2}{1-\psi^2})$ which can be simplified to $2\sigma_\varepsilon^2(1+\psi)^{-1}$. Note also that $p_t - p_{t-1}$ follows an ARMA(1,1) process.

[11] See also Da and Shive (2013) who note that "at least some ETF-driven return co-movement is excessive." Similarly, Broman (2013) argues that "ETF mispricing" is only partially mean-reverting. 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, e.g., Wurgler (2010) and Ramaswamy (2010).

[12] Source: BlackRock. ETF assets are sourced using shares outstanding and net asset values from Bloomberg. Asset classifications are assigned by the BlackRock based on product definitions from provider websites and product prospectuses. We excluded 26 funds for which missing data led to non-convergence of the maximum likelihood procedure.

[13] We also estimated logistic regressions as the dependent variables have limited range. The results are very similar to the ordinary least-squares estimates and are not reported here.

[14] The model is estimated (consistent with our framework) with log values, but for illustrative purposes we plot the dollar values in Figure 1.

[15] This is not just a result of the crisis period. We obtain similar results using a sample from April 13, 2007 to December 31, 2013. The autocorrelation in total NAV daily returns in this period is 0.544 (t-value of 26.65) versus -0.009 (t-value of -0.37) for ETF daily returns.

References

Ben-David, I., Franzoni, F., and Moussawi, R. (2014). "Do ETFs Increase Volatility?," Dice Center WP 2011–20, Ohio State University.

Broman, M. S. (2013). "An Index-Based Measure of Liquidity," Working Paper, Santa Clara University.

Chacko, G., Das, S., and Fan, R. (2013). "Excess Co-Movement and Limits-to-Arbitrage: Evidence from Exchange-Traded Funds," Working Paper, York University.

Da, Z. and Sophie, S. (2013). "Exchange-Traded Funds and Equity Return Correlations," Working Paper, University of Notre Dame.

Dieterich, C. and Cui, C. (2014). "Tuesday's Bond Sell-off Came with Massive ETF Redemptions," *Wall Street Journal*, March 5, 2014.

Engle, R. F. and Sarkar, D. (2006). "Premiums-Discounts and Exchange Traded Funds," *Journal of Derivatives*, Summer **13**(4), 27–45.

Gastineau, G. (2002). *The Exchange-Traded Funds Manual*, John Wiley & Sons.

Golub, B., Novick, B., Madhavan, A., Shapiro, I., Walters, K., and Ferconi, M. (2013). *Viewpoint: Exchange Traded Products: Overview, Benefits and Myths*, BlackRock Investment Institute.

Grégoire, V. (2013). "Do Mutual Fund Managers Adjust NAV for Stale Prices?," Working Paper, University of British Columbia.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press.

Hasbrouck, J. (2003). "Intraday Price Formation in US Equity Index Markets," *Journal of Finance* **58**(6), 2375–2399.

International Monetary Fund (2014). "Global Financial Stability Report: Moving from Liquidity- to Growth-Driven Markets," Washington, DC (April 2014).

Office of Financial Research (2013). "Asset Management and Financial Stability," US Department of Treasury, Washington, DC. Available at: http://financialresearch.gov/reports/files/ofr_asset_management_and_financial_stability.pdf.

Petajisto, A. (2013). "Inefficiencies in the Pricing of Exchange-Traded Funds," Working Paper, New York University.

Poterba, J. and Summers, L. (1988). "Mean Reversion in Stock Prices," *Journal of Financial Economics* **22**, 27–59.

Ramaswamy, S. (2010). "Market Structures and Systemic Risks of Exchange-Traded Funds," Bank of International Settlements, BIS Working Paper No. 343.

Sullivan, R. and Xiong, J. X. (2012). "How Index Trading Increases Market Vulnerability," *Financial Analysts Journal* **68**(2), 70–85.

Tucker, M. and Laipply, S. (2013). "Bond Market Price Discovery: Clarity Through the Lens of an Exchange," *Journal of Portfolio Management* **39**(2), Winter.

Wimbish, W. (2013). "Serious Health Warnings Needed for Some ETFs," *Financial Times*, June 23.

Wurgler, J. (2010). "On the Economic Consequences of Index-Linked Investing," NBER Working Paper 16376, National Bureau of Economic Research, Cambridge, MA.

Keywords: ETF; premiums; discounts; price discovery; arbitrage; liquidity