ETF Arbitrage and Return Predictability*

David C. Brown
Eller College of Management
University of Arizona
dcbrown@email.arizona.edu

Shaun William Davies
Leeds School of Business
University of Colorado, Boulder
shaun.davies@colorado.edu

Matthew C. Ringgenberg
David Eccles School of Business
University of Utah
Matthew.Ringgenberg@Eccles.Utah.edu

First Draft: July 1, 2016
This Draft: December 20, 2017

ABSTRACT

Demand shocks generate mispricing between identical securities leading to arbitrage activity that restores *relative* price efficiency. However, relative price efficiency does not imply *absolute* price efficiency as either demand shocks or subsequent trades by arbitrageurs may push assets prices away from latent fundamental values. In theory, if arbitrage trades are observable, such distortions are short-lived. We examine a novel setting in which arbitrage trades are publicly observable shortly after they occur, and find that market participants *do not* fully incorporate that information into market prices. Instead, arbitrage activity negatively predicts subsequent returns and ETF investors, in aggregate, underperform benchmarks.

Keywords: Arbitrage, Law of One Price, Return Predictability, Exchange Traded Funds, ETFs

JEL Classification Numbers: G12, G14

*We are grateful for the helpful comments of Jesse Blocher, Scott Cederburg, Ed Van Wesep, Nicholas Hirschey, Alex Chinco, Don Chance, Harry Turtle, and conference and seminar participants at Research Affiliates, Cubist Systematic Strategies, the 2016 Brevan Howard Centre for Financial Analysis' Finance Theory Group Conference, HKU, HKUST, University of Colorado, the European Winter Finance Conference 2017, the ASU Sonoran Winter Finance Conference 2017, the 2017 FSU SunTrust Beach Conference, and the 2017 Front Range Finance Seminar. All errors are our own.

1 Introduction

Arbitrageurs play a crucial role in asset pricing. Strictly speaking, arbitrageurs enforce the law of one price, maintaining price efficiency between identical securities. When a demand shock generates a mispricing, arbitrageurs must buy and sell assets to converge the price of the two securities. However, this does necessarily imply that the price converges to the securities' shared fundamental value, i.e., *relative* price efficiency does not equate to *absolute* price efficiency. If arbitrage activity is observable, such distortions are likely to be short-lived as market participants account for the information (or lack thereof) contained in the implicit demand shocks and the arbitrage trades themselves. However, arbitrage activity is difficult to observe in real world data. In fact, the existence of arbitrageurs generates a paradox: if arbitrageurs are successful at correcting mispricings, then there may be no evidence of arbitrage opportunities in the data. Conversely, if the empiricist observes an apparent arbitrage opportunity, it may not be exploitable due to limits to arbitrage. Thus, there are relatively few empirical studies which quantify the trades of arbitrageurs, and a number of important questions remain unanswered. If one could observe the trades of arbitrageurs, would this allow market participants to jointly enforce relative and absolute price efficiency? Put differently, is there information in the trades of arbitrageurs? Moreover, how quickly do other market participants trade on this information?

In this paper, we answer these questions in a setting in which arbitrage trades are observable to all market participants shortly after they occur. We study the exchange-traded fund (ETF) market, in which share creation/redemption activity provides the direction and magnitude of arbitrage activity. When a premium (discount) exists between an ETF's price and its net asset value (NAV), arbitrageurs can simultaneously sell (buy) the ETF and buy (sell) the underlying assets. To complete the transaction, arbitrageurs exchange the underlying assets for newly created ETF shares (or vice versa) in the primary market. Thus, an

ETF's change in shares outstanding measures net arbitrage activity. Importantly, shares outstanding are published on a daily basis, making ETF arbitrage activity observable to all market participants. As a result, weak-form market efficiency implies that any information in arbitrage trades should be quickly incorporated into prices, and that prices' deviations from fundamental value should be short-lived. Our main finding is that market participants do not fully incorporate observable arbitrage trades into prices, and instead, ETF arbitrage activity predicts subsequent asset returns.

Our tests reject weak-form market efficiency, as we show that conditioning on publicly observable arbitrage data yields excess returns that cannot be explained by canonical risk factors. While arbitrage activity is our conditioning variable, we cannot distinguish whether fundamental price distortions arise from (i) arbitrage activity itself (i.e., the buying and selling of mispriced claims), (ii) implicit demand shocks that generate arbitrage opportunities, or (iii) a combination of (i) and (ii). To see this, consider the example in Figure 1 which depicts the values of an ETF share and the share's underlying NAV at three different points in time. At $t = 0$, a small premium exists between the ETF share price and NAV, but it is not large enough to attract arbitrageurs due to transaction costs. At $t = 1$, a demand shock hits both the ETF shares and the underlying assets, but to *different* degrees, leading to a larger mispricing. At $t = 2$, arbitrageurs exploit the mispricing to the point at which the premium is no longer large enough to attract arbitrage trades. Notably, the ETF and NAV prices are pushed by *both* the latent demand shock and arbitrage trades. As such, our aim is not to attribute price distortions squarely on the shoulders of arbitrageurs, nor is it to make welfare claims about the ETF mechanism. Instead, we show that market participants are dismissing valuable public information.

Our analysis of ETF arbitrage activity and return predictability uses data from January 2007 through December 2016 for 2,196 U.S. traded ETFs. We begin by analyzing the effects of arbitrage activity on individual stocks held by ETFs. Ben-David, Franzoni, and Moussawi

(2017a) show that daily ETF share creation and redemption activity is associated with short-term price distortions that reverse over time. Following their methodology, we aggregate ETF share creation and redemption activity at the stock-month level using the portion of assets that each ETF holds in a given stock. We then sort stocks into creation/redemption deciles and examine future stock returns. Univariate sorts show that monthly ETF arbitrage activity generates significant negative return predictability, consistent with Ben-David et al. (2017a), who find that ETF distortions are likely caused by non-fundamental trader demand. A trading strategy that buys stocks held by ETFs with extreme outflows and sells stocks held by ETFs with extreme inflows generates a statistically significant four-factor alpha of 7% per annum. Moreover, these price distortions are stronger in stocks that are less actively traded; a trading strategy that conditions on stocks with low volume earns a slightly larger four-factor alpha of 8% per annum.

Many stocks are included in a multitude of ETFs, so our stock-level results necessarily reflect the aggregation of creations and redemptions across many ETFs. As such, it is not clear that the predictability we find at the individual stock level will translate to ETF prices more generally. In particular, as arbitrage activity is observable at the ETF level, one may expect traders to condition on a particular ETF's share creations/redemptions and adjust prices so that ETFs are, on average, priced efficiently (even if stocks within the ETF portfolio are not). However, we find *stronger* return predictability at the ETF level. Figure 2 shows an event time graph of the return predictability induced by trading in ETFs. We define the event date (t=0) as months with top decile creations or redemptions. As money flows into (out of) ETFs, cumulative abnormal returns rise (fall), creating a return difference of nearly 1%. However, the return gap quickly reverses in the following month. The figure suggests that prices rise (fall) as investors flow into (out of) ETFs, but this price increase (decrease) represents a temporary dislocation that will predictably reverse over one to three months.

To formalize this result, we sort ETFs into deciles based on monthly ETF arbitrage activity. As ETFs are diversified portfolios, we analyze raw returns as well as abnormal returns. Univariate sorts document statistically significant abnormal returns in the range of 11% to 26% per year. We also examine the relation between ETF arbitrage activity and future ETF returns using a regression framework that allows us to control for a variety of fund-level and macro-level characteristics. We again find that ETF creation activity is associated with predictably lower future returns. Low redemption ETFs outperform high creation ETFs by 20% per year. The results reject weak-form market efficiency and, moreover, show that the ETF shares are more sensitive to non-fundamental demand than the underlying assets (if shares were instead *less* sensitive than the underlying assets, arbitrage activity would be associated with return continuations rather than reversals).

Market participants fail to incorporate the information contained in observable arbitrate activity. As such, we expect larger distortions in markets in which investors are ignoring more information, i.e., ETF primary markets with more arbitrage activity. We find this to be the case. ETFs with the most primary market activity, as measured by days of creations or redemptions, show much stronger return predictability. Additionally, return predictability is not driven by more illiquid ETFs that are subject to price non-synchronicity, such as bond or international ETFs.

Our cross-sectional results suggest that ETF investors collectively increase and decrease their exposure to risky assets in a systematic manner that is inversely related to future returns. To quantify the impact of mispricing for a representative investor, we calculate ETF returns on a share-growth-adjusted basis rather than using single-share returns. Share-growth-adjusted returns are essentially asset-weighted, scaling-up returns after net inflows, and scaling-down returns after net outflows. We find that share-growth-adjusted returns are negatively skewed, confirming ETF investors' poor collective timing. As an example, we find that the representative investor in SPY, the largest ETF which accounts for almost 10% of the value in all ETFs, underperformed by 145 basis points per annum. Moreover, share-growth-adjusted returns for the aggregate ETF market suggest underperformance of 7 bps to 33 bps annually. To put this in perspective, the effective underperformance in 2016 amounts to between \$1.6B and \$7.7B. Put differently, the combination of non-fundamental demand shocks and price pressure from arbitrage trades leads to a real wealth loss for ETF investors as they underperform intended benchmark indices.

Our paper makes several contributions to the literature. First, our paper provides new insights into the relation between arbitrage and market efficiency. Because most studies focus on observable mispricings, the existing literature has largely focused on the shortcomings of arbitrage activity (e.g., Pontiff (1996), Shleifer and Summers (1990), Shleifer and Vishny (1997), and Lamont and Thaler (2003)). We add new insights to this literature by focusing on observable arbitrage activity, rather than the absence of it. Specifically, we show that short-term *relative* price efficiency does not necessarily imply longer-term *absolute* price efficiency. While Ben-David et al. (2017a) shows that ETFs can lead to short-term price inefficiencies for individual stocks, we show that ETF creations and redemptions lead to monthly return predictability for both the underlying assets and the ETFs themselves. This is particularly surprising because ETF creations and redemptions are observable by all market participants.[1] Thus, we document that one form of market *inefficiency* is related to the maintenance of another form of market *efficiency*.

Second, we contribute to the growing literature on the relation between ETFs and other market outcomes. A number of papers study the direct effects of ETF arbitrage on assets. Baltussen, van Bekkum, and Da (2016), Da and Shive (2016) show that ETFs induce co-movement between underlying assets, and Ben-David et al. (2017a) and Krause, Ehsani, and Lien (2013) document volatility transmission from ETFs to the funds' underlying assets. We complement these studies by showing that investors collectively ignore ETF arbitrage activ-

[1] Bessembinder (2015) also argues that predictable ETF order flows should have minimal effects on long-term prices.

ity and asset prices slowly revert back to fundamentals following non-fundamental demand shocks. Furthermore, we provide novel evidence that ETF returns and share changes are not independent. Future research examining ETF returns may need to control for both changes in prices (as typical stock return studies do) *and* changes in quantities.

Third, our analysis also contributes to a recently revived discussion of the "arithmetic of active management" (Sharpe, 1991) — the idea that active asset management as a whole must earn zero excess returns before fees. Sharpe's point is simple: passive market indexers earn the market return (by definition), implying that any excess returns generated by active managers must come from other active managers. Thus, active managers, in total, must earn zero excess returns. While intuitive, the notion has come under recent scrutiny. Petajisto (2011) notes that passive index investing strategies will systematically under perform because they miss the positive performance associated with the *announcement* that a stock will be added to the index.[2] This underperformance by passive investors provides the opportunity for overperformance by active management, violating the arithmetic. More recently, Pedersen (2016) outlines additional violations of the arithmetic due to differential market access (e.g., passive investors' inability to participate in IPOs) and trading costs due to rebalancing. Our share-growth-adjusted return results suggest another significant way in which Sharpe's arithmetic may not hold. We show that ETFs tend to underperform their target index, i.e., an increase (decrease) in shares outstanding is subsequently accompanied with underperformance (overperformance). Thus, this underperformance provides an additional source of excess returns for active management.

[2]There is an extensive literature documenting positive abnormal returns after the announcement that a stock will be included in a major stock index, e.g., see Shleifer (1986).

2 ETF and Sample Details

The U.S. ETF market has grown dramatically over the last decade; total ETF assets have gone from $151 billion in 2003 to over $3 trillion in 2016 (BlackRock (2014) and Madhavan (2016a)). Accordingly, academics, practitioners, and regulators, have all become increasingly interested in the structure of the ETF market and its impact on financial markets. In this section, we provide an overview of the institutional details and existing academic research regarding ETFs.[3]

Like mutual funds, ETFs are pooled investment vehicles which allow investors to buy a basket of assets in the secondary market.However, unlike a mutual fund, shares of ETFs can be created or redeemed in the primary market. ETF sponsors (e.g. iShares and State Street) create a primary market by publishing the baskets of securities that may be exchanged for ETF shares (or vice versa), and by designating authorized participants (APs, who are mostly large institutional investors), who can transact on the primary market. The primary market is designed to equilibrate supply and demand for shares in the ETF, and allows APs to effectively enforce the law of one price in real time.

For example, suppose a non-fundamental demand shock hits an ETF, pushing its price above the NAV of the underlying stocks. The AP can then simultaneously transact in the secondary markets, selling short the ETF and buying the underlying basket of stocks, locking in the profitable price premium.[4] The AP can complete the arbitrage by transacting with the ETF sponsor, exchanging the underlying basket of stocks for new ETF shares, covering the original short position.[5] In this example, secondary market trading puts downward

[3]In the interest of brevity, we omit a comprehensive discussion of ETFs and the related literature. For more details, see Madhavan (2016b) and Ben-David, Franzoni, and Moussawi (2017b).

[4]While only APs can transact in an ETF's primary market, giving them a cost advantage, other market participants can conduct similar arbitrage trades, e.g. high-frequency traders, hedge funds and statistical arbitrage traders, among others.

[5]Primary market transactions typically take place at the end of the day and are only conducted in set sizes, known as creation units, which are mostly commonly 50,000 shares. Primary market transactions typically costs $500 to $3,000, regardless of the number of creation units.

pressure on the ETF price and upward pressure on the underlying stock prices, reducing the price premium. Through a similar process, APs can arbitrage supply shocks to ETFs by simultaneously selling the underlying assets and buying ETF shares, and then redeeming the ETF shares for the underlying assets. By continuously conducting such arbitrage trades, APs enforce relative price efficiency between ETFs and their underlying assets.[6]

Because ETF arbitrage involves trades in both the ETF and the underlying assets, a number of papers argue that trading in ETFs can impact the properties of the underlying assets in the ETF portfolio. Specifically, several papers argue that ETFs can change the correlation structure of stock returns. Da and Shive (2016) and Staer and Sottile (2016) show that ETF arbitrage can lead to comovement in equity returns. Similarly, Baltussen et al. (2016) show that serial correlation in equity returns goes from positive to negative after the introduction of ETFs around the world. Moreover, a number of papers argue that ETF trading can allow shocks to be transmitted to the underlying assets. Israeli, Lee, and Sridharan (2015) present empirical evidence that stocks with greater ETF ownership experience relatively worse price efficiency.[7] In contrast, Glosten, Nallareddy, and Zou (2016) examine net changes in ETF positions and they document increased price efficiency for ETF-owned stocks. In addition, Krause et al. (2013) argue that ETFs transmit volatility to stocks. Stocks that are owned by ETFs experience increased volatility resulting from demand shocks (Ben-David et al., 2017a), and increased liquidity commonality (Agarwal, Hanouna, Moussawi, & Stahel, 2017). Consistent with this, Coles, Heath, and Ringgenberg (2017) find that index investing is associated with higher volatility, higher comovement in stock

[6] In general, pricing differences are small, but they can be time-varying and some can be persistent. Petajisto (2017) shows that the average difference between ETFs and their net asset values is only 6 basis points, but that the volatility of the difference is 49 basis points, suggesting substantial variation across ETFs and across time. Engle and Sarkar (2006) shows that the premiums (discounts) between ETFs and their underlying assets are generally very small, and when they do exist, they last only a few minutes. In contrast, Fulkerson and Jordan (2013) finds ETF premiums and discounts can persist for five days.

[7] In related work, Staer (2016) documents contemporaneous price pressure on the underlying stocks held by ETFs and subsequent reversals.

returns, and prices that are less likely to follow a random walk. These results support several theoretical models. Specifically, A. Bhattacharya and O'Hara (2016) show that ETFs can allow non-fundamental shocks to propagate into underlying asset prices. Similarly, Malamud (2015) shows theoretically that the ETF arbitrage mechanism may lead to higher volatility and momentum in the prices of underlying assets. However, the model also shows that the introduction of new ETFs may actually result in a reduction of volatility due to a demand substitution effect.

As baskets of stocks, ETFs are naturally compared to closed-end funds and index funds, but important distinctions make ETFs uniquely suited for studying arbitrage. For instance, closed-end funds and ETFs both trade on secondary markets, and have primary markets. However, closed-end funds are not transparent, rarely issue or redeem shares, and any transactions are at the discretion of fund sponsors. Alternatively, ETFs' primary markets are regularly open to all APs and holdings are published daily. An additional important distinction is that maintaining price efficiency is decentralized for ETFs. ETFs are similar to index funds in that they are both subject to daily investor flows. However, index fund managers have discretion over how to invest or divest to manage these flows, potentially doing so in a manner that reduces price impact of trading.[8] In contrast, the decentralized nature of ETFs encourages arbitrage activity, resulting in the transmission of flow shocks into the prices of ETFs and their underlying assets.

2.1 Data

To examine the relation between arbitrage activity and asset prices we combine data from Bloomberg, CRSP, and Kenneth French's website. From Bloomberg, we get daily data on

[8] More generally, active and passive managers can actively manage funds flows, which can have an impact on underlying asset prices (e.g., Coval and Stafford (2007), Lou (2012), Cella, Ellul, and Giannetti (2013), Hombert and Thesmar (2014), Arif, Ben-Rephael, and Lee (2016), and Huang, Ringgenberg, and Zhang (2017)).

ETF share prices, ETF NAVs, ETF shares outstanding, and ETF trading volumes.[9] Each date, we calculate ETF premiums (discounts) as the difference between each ETF's price and its NAV. We then merge this data with information from CRSP including Lipper Codes and stock returns, as well as holdings data for many of our sample ETFs. In our stock level analyses, we limit our data to those stocks with a price greater than $5 and a CRSP share code of 10 or 11. Finally, to calculate a risk-adjusted measure of returns, we add information on the three-factor (Fama & French, 1993), four-factor (Carhart, 1997), and five-factors models (Hou, Xue, and Zhang (2016), Fama and French (2015)) from Kenneth French's website.

Table 1 displays a time-series count of the number of ETFs in our sample. As previously discussed, the ETF market has grown rapidly over the last decade, and by the end of 2016, our sample includes 1,707 unique ETFs. To mitigate the impact of illiquidity and possible non-synchronous prices due to infrequent trading, we limit our sample to ETFs with at least $50 million in assets. 31% of ETFs are excluded using the $50 million threshold, but they collectively account for less than 1% of market capitalization. We also consider a sample of ETFs that are flagged as "mature" once they exceed the $50 million threshold and experience a month in which at least one-half of the trading days had some share creation/redemption activity. This is to ensure that we are analyzing ETFs with active primary and secondary markets. As shown in Table 1, this filter removes approximately half of the remaining ETFs, but only reduces the total market capitalization by 9%.

Table 2 displays summary statistics for the sample of $50M+ ETFs, as well as the mature sample of ETFs.[10] As expected, the mature sample is larger and generally experiences more

[9] A number of ETFs have anomalous data on prices and shares outstanding that appear to be incorrect. Rather than winsorizing our data, we clean the data by removing the anomalies that are not verifiable via other data sources. See the data appendix for more details on database construction and cleaning. Furthermore, Ben-David et al. (2017a) suggest that Bloomberg provides the most accurate daily ETF data.

[10] The entire sample of ETFs is omitted for the sake of brevity as the $50M+ sample is representative of the entire sample.

trading and better liquidity; mature ETFs have more shares outstanding, more turnover, and tighter bid-ask spreads. In Panel B of Table 2, we display information on the Lipper Categories of the ETFs. While the two samples are fairly similar, the mature ETFs tend to be more focused on equities and less focused on more exotic asset classes like bonds and international equities.

In the tests that follow, we examine the relation between returns and creation/redemption activity aggregated to the month-level. We avoid higher frequency measures for several reasons. First, the accounting standards for share creation/redemption activity vary across ETFs — some funds use $T+1$ accounting (i.e., they register the share creation activity the day after it occurs) while other funds use T accounting. Moreover, these accounting standards have changed over time, and the change from $T+1$ to T accounting, or vice versa, is not public.[11] Second, there is some evidence that authorized participants may strategically delay creating or redeeming shares to take advantage of failure-to-deliver rules at clearing houses. Evans, Moussawi, Pagano, and Sedunov (2016) describes how authorized participants can wait until $T+6$ to create new shares and thus avoid costs associated with short-selling. Accordingly, by focusing on monthly returns and shares outstanding we mitigate the impact of these effects, and thus the predictability we identify is unlikely to be due to microstructure effects or institutional details related to the creation/redemption mechanism. Furthermore, using monthly data provides ample time for market participants to incorporate arbitrage activity into prices, making it more likely that weak-form market efficiency holds.

3 Arbitrage Activity and Price Efficiency

While we are interested examining the relation between arbitrage activity and *absolute* price efficiency, several existing papers examine the relation between arbitrage activity and *relative*

[11]See Staer (2016) for additional details.

price efficiency. As previously discussed, Engle and Sarkar (2006), Fulkerson and Jordan (2013), and Petajisto (2017) examine deviations between ETF share prices and the value of the ETFs' underlying assets. In general, all three papers find that deviations of the law of one price are relatively small and when they get larger, they are corrected relatively quickly. In other words, arbitrageurs in the ETF market are able to ensure relative price efficiency. Accordingly, we take as given that authorized participants act to correct violations of the law of one price. Our paper is unique in that we examine whether market participants incorporate observable arbitrage activity into asset prices so that absolute price efficiency also holds.

3.1 Stock Returns

We begin by examining the relation between ETF arbitrage activity and the absolute price efficiency of the individual stocks held by ETFs. To measure arbitrage activity, we calculate creation and/or redemption activity in a given ETF. Formally, we define ETF arbitrage activity as the percentage change in ETF shares outstanding for fund j at time t:

$$ETFArb_{j,t} = \frac{SharesOutstanding_{j,t}}{SharesOutstanding_{j,t-1}} - 1. \quad (1)$$

We convert the percentage change in ETF shares outstanding into a dollar flow $ETFFlow_{j,t}$ by multiplying $ETFArb_{j,t}$ by ETF j's period t market capitalization $ETFMarketCap_{j,t}$,[12]

$$ETFFlow_{j,t} = ETFArb_{j,t} \times ETFMarketCap_{j,t}. \quad (2)$$

Intuitively, $ETFArb_{j,t}$ measures arbitrage activity in a fund as a percent of shares outstanding, while $ETFFlow_{j,t}$ measures the dollar amount of arbitrage activity.

[12]Our dollar flow measure is similar to that used in e the mutual fund literature (e.g., Coval and Stafford (2007))

We aggregate ETF-level dollar flows to a stock-level measure of arbitrage activity using each stock i's portfolio weight in each ETF j. Specifically, each $ETFFlow_{j,t}$ is pro-rated to the stock-level, summed across all J ETFs, and then normalized by the stock's market cap,

$$StockArb_{i,t} = \frac{\sum_{j=1}^{J} ETFFlow_{j,t} \times Weight_{i,j,t}}{MarketCap_{i,t}}, \tag{3}$$

where $ETFFlow_{j,t}$ is the dollar amount of creations or redemptions in ETF j on date t, $Weight_{i,j,t}$ is the percentage of ETF j held in stock i on date t and $MarketCap_{i,t}$ is the market capitalization of stock i on date t. The resulting variable, $StockArb_{i,t}$, is the normalized arbitrage dollar flow in a particular stock.

Using this measure, we examine the relation between arbitrage activity and absolute price efficiency at the stock-level. We start by examining portfolio sorts. Each period, we sort stocks into deciles based on $StockArb_{i,t}$. Portfolio one contains stocks with the highest arbitrage outflows and portfolio ten contains stocks with the highest arbitrage inflows. We also use dual sorts that first sort on firm-level measures of liquidity (by terciles) and then on our measure of ETF arbitrage activity.

The results are shown in Table 3. The table displays one month ahead excess returns and portfolio alphas (in percent) calculated using the four-factor model. In both panels, the results are clear: arbitrage activity is associate with substantial mispricing in stocks, especially those with low volume. In Panel A, *All Stocks*, we find that a trading strategy that buys stocks held by ETFs with extreme outflows and sells stocks held by ETFs with extreme inflows generates a statistically significant excess return of 61 bps per month (7.3% per annum) and a statistically significant four-factor alpha of 60 bps per monh (7.2% per annum). In Panels B and C, we then split the sample on measures of stock-level liquidity. Specifically, Panel B examines the returns after sorting stocks into terciles based on bid-ask spread, while Panel C examines the returns after sorting stocks into terciles based on volume as a percent of market capitalization. In both cases, we use conditional sorts that sort first on the firm specific measure of liquidity and then sort on our measure of arbitrage activity. In Panel B, the results are approximately half the size of the main effect; in other words, bid-ask spread as a measure of liquidity seems to explain a significant portion of the price effect from ETF arbitrage. However, while the bid-ask results are not statistically significant, we note that these tests have significantly lower power than the results in Panel A. Interestingly, in Panel C, we continue to find evidence of significant return predictability even after we condition on stock-level volume. Specifically, we find that the return predictability from ETF arbitrage is highest in stocks with low to medium trading volume. Overall, the portfolio sorts suggest that (i) ETF share redemption activity is associated with a non-fundamental negative demand shock that disproportionately affects stocks with low volume; (ii) high selling pressure by authorized participants causes stock prices to fall below fundamental values, or (iii) a combination of (i) and (ii). Regardless of the source for pushing these stock prices *below* their fundamental values, the prices later predictably reverse.

The portfolio sort analyses suggest that firm characteristics, like volume, are important determinants of returns. Accordingly, to better control for time-invariant firm characteristics, as well as macro-economic and industry trends, we turn to a regression setting with fixed effects. Formally, we examine OLS panel regressions of the form:

$$Ret_{i,t+1} = \beta_1(StockArb_{i,t}) + FE_i + FE_{j\times y} + \epsilon_{i,t+1}, \tag{4}$$

where $Ret_{i,t+1}$ is the one-month ahead excess stock return (in percent) from CRSP, $StockArb_{i,t}$ is the price pressure in stock i in month t due to ETF arbitrage activity, FE_i is firm fixed effects, and $FE_{j\times y}$ is industry $\times$ year fixed effects. The firm fixed effects allow us to account for firm characteristics (like size and liquidity) while the industry-year fixed effects control for time-varying changes to industry conditions as well as general macroeconomic trends.

The results are shown in Table 4 with t-statistics calculated using standard errors cluster by firm and year-month shown below the estimates. The regression results confirm the portfolio sort results: we find that arbitrage dollar flows are associated with significant return predictability at the stock level. In column (2), which includes firm and industry-year fixed effects, the statistically significant coefficient estimate of -0.013 implies that a one-standard deviation increase in $StockArb_{i,t}$ is associated with a 27 basis point decrease in one-month ahead returns at the stock level (approximately 3.2% per annum).[13]

In columns (5) through (8), we test for a non-linear relation between arbitrage activity and future stock returns by ranking stocks into deciles based on $StockArb_{i,t}$. We include an indicator variable in the regression equation to denote membership in a particular decile, with the lowest decile ($StockArb_{i,t} = 1$) as the omitted category. The results are concentrated in deciles 8 through 10 (stocks that experienced large purchases by authorized participants), suggesting a convex relation between arbitrage activity and subsequent returns. In column (6), which includes firm and industry-year fixed effects, the statistically significant coefficient estimate of -0.697 on Decile 10 implies that stocks with large purchases by authorized participants underperform stocks with largest sales by authorized participants (the omitted case) by 69 basis points over the next month (approximately 8.2% per annum).

Overall, our findings suggest that ETF arbitrage leads to return predictability at the stock level. Thus, while arbitrage activity induces *relative* price efficiency, market participants appear to be neglecting the information contained in this observable arbitrage activity.

3.2 ETF Returns

As baskets of underlying assets, ETFs will naturally aggregate any return predictability in those underlying assets, suggesting that ETF arbitrage activity also predicts ETF returns. However, as assets vary in the degree of return predictability related to arbitrage activity, and

[13] The standard deviation of $StockArb_{i,t}$ is 20.85; -27 bps = -0.013 $\times$ 20.85.

as the correlation in return predictability among the underlying assets is unknown, return predictability in an ETF's underlying assets does not necessarily imply return predictability for the ETF itself. Moreover, as arbitrage activity is observable at the ETF level, one may expect traders to condition on a particular ETF's share creations/redemptions and adjust prices so that ETFs are, on average, priced efficiently (even if a few stocks within the ETF portfolio are not). In this section, we test whether or not arbitrage activity predicts future ETF returns.

We start by sorting ETFs into portfolios based on last month's $ETFArb_{j,t}$, which measures share creation or redemption in ETF j at time t. In all of our portfolio sorts, we sort based on characteristics at a monthly level, preventing time trends and differences in sample size from driving our results. Our portfolio sorts are designed to test whether past ETF arbitrage activity by authorized participants is related to future ETF performance. We measure ETF performance using ETF returns from the month following portfolio formation.[14] ETFs with the most creation activity in the past month are sorted into Decile 10, and ETFs with the most redemption activity are sorted into Decile 1.

Panel A of Table 5 displays the results for both raw returns and four-factor abnormal returns.[15] The table shows equal-weighted and value-weighted returns in both the \$50M+ and the mature ETF samples. Based on raw returns, Panel A of Table 5 shows that ETFs that have experienced large creation activity underperform ETFs that have experience large redemption activity. The difference in monthly raw returns between Decile 1 and Decile 10 range from 83 basis points to 199 basis points (10.4% to 26.7% annualized). The differences are larger and more significant using equal-weighted portfolio returns (t-statistics of 3.223

[14] Changes in ETF premia over NAV create a small discrepancy between ETF returns and NAV returns. However, the volatility of ETF premia are relatively small compared to the volatility of the basket of underlying assets. Because we focus on a monthly horizon, our results are qualitatively similar using either NAV returns or ETF returns.

[15] Results are qualitatively similar using a three-factor or five-factor model. The choice of factor model does not qualitatively change any of our results, and we only present four-factor results going forward.

and 4.010 for the $50M+ and mature samples). Using value-weighted portfolio returns, the return differences are smaller and t-statistics and less significant (2.042 and 2.203). In both samples, the results are stronger for mature ETFs, suggesting that our results are not driven by new, relatively illiquid ETFs. Incorporating factor returns does not significantly change the results, as the test portfolios combine well-diversified ETFs. The differences in alpha estimates are consistent with the results using raw returns, ranging from 86 basis points to 214 basis points (10.8% to 28.9% annualized), and statistical significance is also similar. Overall, our portfolio sorts provide evidence of strong return predictability based on past ETF share creations and redemptions.

To examine the robustness of our portfolio sorts, we next examine a panel regression of the form:

$$Ret_{j,t+1} = \alpha + \sum_{d=1:10} \beta_d Decile_{j,d,t} + \Gamma X_t + \delta V_{j,t} + \alpha_j + \epsilon_{j,t+1}, \quad (5)$$

where $Ret_{j,t+1}$ is next month's return on ETF j (including distributions), $Decile_{j,d,t}$ is an indicator variable for whether ETF j is in decile portfolio d in period t, X_t are factor returns, $V_{j,t}$ are ETF characteristics, and α_j are ETF fixed effects. As before, we calculate robust standard errors clustered by ETF and year-month. Columns (1)–(3) in Table 6 show the results of estimating Equation 5. All specifications include fixed effects, the second and third columns include contemporaneous four-factor returns and the third column includes lagged ETF returns, lagged premia, lagged market capitalizations and lagged volumes. Consistent with our prior results, there are significant differences in return predictability between ETFs in the Decile 1 and Decile 10 portfolios. Across the three columns, monthly returns are 154 to 158 basis points (about 20% annualized) lower for Decile 10 ETFs relative to Decile 1 ETFs (the omitted group). Deciles 2 through 9 also have lower returns than Decile 1, but the differences are smaller (ranging from 26 to 65 basis points) and are only sometimes significant. A lack of a clear pattern across Deciles 2 through 9 suggests that most of the return predictability occurs in ETFs with high creation or redemption activity, and thus the relation between past share changes and future returns may be non-linear.

To directly test for a linear relation, we replace the deciles in Equation 5 with a continuous measure of arbitrage activity:

$$Ret_{j,t+1} = \alpha + \beta ETFArb_{j,t} + \Gamma X_t + \delta V_{j,t} + \alpha_j + \epsilon_{j,t+1}. \tag{6}$$

Columns (4)–(6) show that the point estimates on the coefficients are negative, but the results are not statistically significant, confirming the non-linear relation between past share changes and future returns. As a final result in this section, Columns (7)–(8) split our sample period into 2007–2011 and 2012–2016. The results show that significant differences in return predictability appear in both periods, although the economic effect is weaker in the later period (26.1% versus 18.7% annualized).

3.3 Which ETFs drive our results?

Our prior results show that markets fail to incorporate the information contained in observable arbitrate activity. As such, we expect larger distortions in markets in which investors are ignoring more information, i.e., ETF primary markets with more arbitrage activity. Returning to Table 5, Panel B displays results of portfolio sorts after first sorting ETFs into terciles based on primary market activity, as measured by days of either creations or redemptions in the past month (i.e., the number of days with an absolute change in shares outstanding). The results show higher return differences among ETFs with more primary market activity. In the lowest tercile of primary market activity, returns between high creation and high redemption ETFs are not statistically different. In the medium tercile of primary market activity, high redemption ETFs outperform high creation ETFs by between

150 and 192 bps (19.6% and 25.6% annualized), and in the high tercile, the out-performance ranges from 182 to 291 bps (24.2% and 29.7% annualized). These results reinforce the idea that markets ignore the information contained in ETF arbitrage trades.

While our results suggest that the information contained in arbitrage trades are ignored, it is possible that ETFs holding illiquid assets or assets with non-synchronous pricing may be driving our results. To address such concerns, we test whether the relation between ETF arbitrage and future returns is different for different types of ETFs. Specifically, we introduce indicator variables for specific ETF characteristics into a panel regressions of the form:

$$\begin{aligned} Ret_{j,t+1} &= \alpha + \beta_1 Decile1_{j,t} + \beta_2 Decile1_{j,t} \times ETFCharacteristic_j + \beta_3 Decile10_{j,t} \\ &\quad + \beta_4 Decile10_{j,t} \times ETFCharacteristic_j + \Gamma X_t + \alpha_j + \epsilon_{j,t+1}, \end{aligned} \tag{7}$$

where $Ret_{j,t+1}$ is the monthly return on ETF j including distributions, $DecileD_{j,t}$ is an indicator variable for whether ETF j is in decile portfolio D in period t, X_t are factor returns, $ETFCharacteristic$ represent indicator variables for certain ETF characteristics, and α_j are ETF fixed effects. The ETF characteristics we consider include an indicator for levered and inverse ETFs (including 2X, 3X, -1X, -2X and -3X funds) and indicators for broad asset-class categories based on funds' Lipper codes, specifically broad equities, sector equities, bonds, commodities and international assets.[16]

Table 7 displays the results. Column (1) provides a baseline reference based on the estimation from Column (2) of Table 6. Column (2) of Table 7 shows a striking result — levered and inverse ETFs display significantly more return predictability relative to non-levered ETFs. Decile 10 levered-ETFs have returns that are 259 basis points lower, and Decile 1 levered-ETFs have returns that are 155 basis points higher, than their non-levered counterparts. Levered ETFs are unique in several ways, all of which may lead to increased

[16]Mixed-asset ETFs and municipal-bond ETFs are included in the broader bond category.

predictability. First, levered ETFs have disproportionately high primary market activity in our sample, which we have shown is consistent with more predictability. Second, leverage accentuates returns, adding power to our tests. Third, levered and inverse ETFs which are small, niche, and derivatives-based, are designed to give investors magnified exposure to benchmark indices. Given the short-term, betting-like nature of levered ETFs, the arbitrage activity in levered ETFs may proxy for broader, market-wide investor sentiment shocks.[17] While our results are largely driven by levered and inverse ETFs, we also note that unlevered ETFs in Decile 10 continue to underperform unlevered ETFs in Decile 1 by approximately 19bps per month (2.3% per year). In other words, our results are strongest in levered ETFs because they have the most arbitrage activity, but our results exist in unlevered funds too (although the magnitude and statistical signficance of the result is smaller).

The remaining columns of Table 7 show that no particular asset class drives our results. Based on point estimates, bond, commodity, international, and sector-based-equity ETFs all show attenuated return differences following large share creations and redemptions relative to other ETFs. For broad-based equity ETFs, the point estimates suggest a larger difference in returns. While not statistically significant, Decile 1 ETFs have 102 bps higher returns (t-statistic of 1.52), and Decile 10 ETFs have 230 basis points lower returns (t-statistic of 3.32), suggesting a wider gap between extreme decile portfolios for the largest and most established group of ETFs. Importantly, these results indicate that our result is not driven by ETFs composed of small, illiquid or non-synchronous assets, and instead, the effects may exist more broadly in the largest category of ETFs.

In short, the results in this section show that ETF arbitrage activity is not incorporated into asset pricing, neither at the stock nor ETF level. Put differently, our findings show that

[17] Davies (2017) studies the original set of levered ETFs offered to investors and uses those ETFs' arbitrage activity to construct a measure of speculation sentiment. The measure of speculation sentiment is show to predict aggregate market return reversals. (Cheng & Madhavan, 2009) shows that levered ETFs are speculative instruments that are not suitable for buy-and-hold investors.

arbitrage activity enforces *relative* price efficiency, but not *absolute* price efficiency. Importantly, these results imply that index investors may underperform their target benchmarks due to price distortions from ETF arbitrage.

4 Arbitrage induced underperformance in ETFs

In Section 3, we find that ETF arbitrage is associated with subsequent return reversals at both the stock and ETF level. Accordingly, in this section, we examine the implications of the mispricing for a hypothetical investor. To do this, we develop a methodology which we refer to as the *share-growth-adjusted return.* To calculate a share-growth-adjusted return for each ETF, we take its return series $\vec{r} = \{r_1, \ldots, r_T\}$ and its one-period-lagged share growth series $\vec{g} = \{g_0, \ldots, g_{T-1}\}$ and perform the following calculation,

$$\mathbf{R} \equiv \prod_{\tau=1}^{T} \big((1 + r_\tau)(1 + g_{\tau-1}) - g_{\tau-1}\big). \tag{8}$$

$\mathbf{R}$ in the preceding calculation is a pseudo portfolio return over the sample period — the analytic expression of $\mathbf{R}$ captures the notion that share creations and redemptions have a leverage-like effect on a fund's total return. For example, if all ETF shares were collectively held by a representative investor, share creations would provide the investor with greater exposure to the ETF's benchmark and redemptions would reduce exposure. We convert $\mathbf{R}$ into an annual return according to:

$$(1 + \mathbf{r}) = \mathbf{R}^{12/T}, \tag{9}$$

and we refer to $\mathbf{r}$ as the *share-growth-adjusted return.* We also define the expected share-growth-adjusted return as,

$$\bar{\mathbf{r}} \equiv E\left[\mathbf{R} | \, \vec{r} \perp \vec{g}\right]^{12/T}, \tag{10}$$

in which the expected share-growth-adjusted return is calculated under the assumption that the time-series of ETF arbitrage activity is orthogonal to the time-series of ETF returns. Expected share-growth-adjusted returns and their underlying distributions are attained via Monte Carlo simulation of one million paths for each ETF. In a given Monte Carlo path k, the vector $\{g_0, \ldots, g_{T-1}\}$ is shuffled into a new vector $\{\hat{g}_{0,i}, \ldots, \hat{g}_{T-1,k}\}$ using the stationary bootstrap technique of Politis and Romano (1994).[18] The stationary bootstrap technique incorporates auto-correlation in the return series and share growth series into our Monte Carlo paths. Using the random vector $\{\hat{g}_{0,k}, \ldots, \hat{g}_{T-1,k}\}$ in Monte Carlo path k, the pseudo portfolio return is calculated as,

$$\mathbf{R}_k^{MC} \equiv \prod_{\tau=t}^{T} \big((1 + r_\tau)(1 + \hat{g}_{\tau-1,k}) - \hat{g}_{\tau-1,k}\big), \tag{11}$$

and its corresponding annual return $\mathbf{r}_k^{MC}$ is calculated as,[19]

$$(1 + \mathbf{r}_k^{MC}) = (\mathbf{R}_k^{MC})^{12/T}. \tag{12}$$

Using the Monte Carlo simulated distribution, we test if an ETF's realized share-growth-adjusted return is statistical different than the expected share-growth-adjusted return, $\mathbf{r} \neq \bar{\mathbf{r}}$. Implicitly, the test is examining if $\vec{g}$ is indeed orthogonal from $\vec{r}$. Moreover, because our stock and ETF level results suggest a negative relation, we perform a one-tail test to examine whether realized share-growth-adjusted returns are *lower* than expected share-growth-adjusted returns, $\mathbf{r} < \bar{\mathbf{r}}$. For completeness, we also perform a one-tail test to examine whether

[18]Unlike a standard bootstrapping method in which observations are chosen randomly with replacement, the stationary bootstrap method picks a random *series* of observations with replacement and the *length* of the series is also random. The stationary bootstrap is a type of block bootstrapping which handles serial correlation in time series data. In our analysis, the length of a drawn series is distributed according to a geometric distribution characterized by $p = \frac{1}{5}$. As such, the average length of a random series is equal to five consecutive observations (i.e., 1/p = 5).

[19]Due to some periods of large share creation changes and some periods of large return swings, some Monte Carlo simulation paths result in $R_k^{MC} < 0$. In these settings, we set $R_k^{MC} = 0$ as an absorbing state, i.e., the ETF goes out of business.

realized share-growth-adjusted returns are *higher* than expected share-growth-adjusted returns, $\mathbf{r} > \bar{\mathbf{r}}$. Under the null hypotheses, ETF realized share-growth-adjusted returns are statistically indistinguishable from the expected share-growth-adjusted returns.

We begin with our sample of mature ETFs and restrict the analysis to 412 ETFs for which we have at least 36 months of data. Table 8 documents a summary of the results at p-value thresholds of 1%, 2.5%, 5%, and 10%. In Panel A, the first column reports the percentage of ETFs, based on equal weights, for which the realized share-growth-adjusted return $\mathbf{r}$ is smaller than thresholds of 1%, 2.5%, 5%, and 10% of the distribution's observations. The second column reports the percentage of ETFs for which the $\mathbf{r}$ is larger than 99%, 97.5%, 95% and 90% of the distribution's observations. Based on the results of Panel A, $\mathbf{r}$ is frequently smaller than what would occur by chance, e.g., 3.16% of the sample falls below the 1% threshold. Furthermore, $\mathbf{r}$ is larger than 99% of observations for only 0.5% of the sample, which is half as frequently as would occur by chance. When ETFs are equally weighted, realized share-growth-adjusted returns are slightly negatively skewed and have a fat left tail. In other words, the results suggest that $\vec{g}$ and $\vec{r}$ are negatively correlated. Thus, investors are more likely to flow into ETFs when they are overpriced and more likely to flow out of ETFs when they are underpriced, and this leads to underperformance.[20]

The second two columns in Panel A report the analysis when ETFs are weighted by assets. We do not simply use end-of-2016 AUM and instead use an asset weight we term *average market capitalization share*: for each fund j, average market capitalization share is computed as,

$$\overline{MktCap}_j = \frac{\sum_{\tau}^{T} \left(\frac{MktCap_{j,\tau}}{\sum_{k}^{J_\tau} MktCap_{k},\tau} \right)}{T}, \tag{13}$$

where the term within parenthesis in (13) represents the fraction of the ETF market that fund

[20]U. Bhattacharya, Loos, Meyer, and Hackethal (2016) studying individual investors' use of ETFs and find evidence of poor timing.

j has in year τ of the sample.[21] The results reported in columns 3 and 4 document dramatic negative skew: over 14% of ETFs' realized share-growth-adjusted returns $\mathbf{r}$ are smaller than the 1% threshold in the simulated distribution. Conversely, only 0.3% of ETFs' realized share-growth-adjusted returns $\mathbf{r}$ are larger than the 99% threshold. At other thresholds, the results are qualitatively the same: realized share-growth-adjusted returns systematically fall below simulated share-growth-adjusted returns at a frequency that is substantially larger than what would happen by chance.

Panel B and Panel C of Table 8 repeat the analysis but over two different subsamples of observations. Panel B repeats the analysis starting in January 2007 and ending in December 2011. Both the equal weighted and market cap share weighted results, in general, are stronger as compared to the results in Panel A. Panel C performs the analysis starting in January 2012 and ending in December 2016. On an equal weighted basis, the results are significantly weaker — the distribution exhibits negative skew, but the tails are not as fat. However, on a market-cap-share-weighted basis, the results remain strong at p-value thresholds of 2.5%, 5%, and 10%. The combination of Panels A, B, and C suggest that ETF arbitrage activity is negatively correlated with subsequent returns in the time series, but the negative correlation has weakened in recent periods.

Our share-growth-adjusted return accounts for the dynamic performance of an ETF given that both prices and quantities (i.e., assets under management) fluctuate. In comparison, a typical return analysis is performed by examining the return on a given share of an ETF that is never redeemed. Focusing on a single share's return does not account for changing ETF (i.e., size). If changes in quantities and changes in prices are unrelated, then examining

[21] We do not use raw market capitalization weights due to non-stationarity in fund sizes during the sample period: the ETF market has been characterized by both rapid growth and the rapid introduction of new funds. As such, weighting by raw market capitalization can be problematic. For example, if one were to use average market capitalization over the sample period, the rapid growth in assets under management would bias the weighting scheme towards new, larger ETFs. Our average market capitalization share controls for this non-stationarity by ranking funds according to their relative share of ETF assets in year τ and the summation takes into consideration only the funds that existed in year τ.

the return on a single share should be equivalent to examining the share-growth-adjusted return. However, our analysis suggests that the changes in quantities and prices are related. Therefore, studies examining ETF returns may need to control for both changes in prices (as typical stock return studies do) *and* changes in quantities.

While our results do not speak to general welfare, our results do have important implications for ETF investors. If all ETF shares were held by a single representative investor, her performance would be closer to that of our share-growth-adjusted return than the performance of a single ETF share. Obviously ETFs are not held by one individual, but allocating returns is a zero sum game. Thus, while a buy-and-hold investor may earn a return consistent with the return of a single share, other higher frequency traders must absorb the residual performance. To put things in perspective, consider the ETF SPY which is State Street Global Advisor's ETF that mimics the S&P500 index via full replication. SPY has a share-growth-adjusted return $\mathbf{r}$ of 5.44% which differs from its simulated expected value $\bar{\mathbf{r}}$ of 6.92% by 148 bps basis points.[22] Thus, while SPY's management fee is 9 bps per annum, our analysis suggests that investors bear an additional indirect cost that is an order of magnitude larger.

Finally, our methodology allows us to quantify the aggregate impact of arbitrage activity on subsequent returns for all ETF investors. Specifically, we take our mature ETF sample and calculate each ETF's monthly arbitrage dollar flow $ETFFlow_{j,t}$ as outlined in Section 3.1. We take the sum of all J ETFs' dollar flows in month t to calculate an aggregate dollar flow,

$$AggFlow_t = \sum_{j=1}^{J} ETFFlow_{j,t}. \tag{14}$$

We then calculate aggregate arbitrage activity as a fraction of aggregate ETF assets in month

[22] As an additional point of reference, the annualized return for a single share of SPY over the sample horizon (which does not account for share creations and redemptions) was 6.89% which is 145 basis points higher than SPY's share-growth-adjusted return.

t as,

$$AggArb_t = \frac{AggFlow_t}{\sum_{j=1}^{J} ETFMarketCap_{j,t}}, \tag{15}$$

where $ETFMarketCap_{j,t}$ is ETF j's market capitalization at the end of period t. We also calculate a market-capitalization-weighted aggregate return,

$$AggRet_t = \frac{\sum_{j=1}^{J} R_{j,t} \times ETFMarketCap_{j,t}}{\sum_{j=1}^{J} ETFMarketCap_{j,t}}. \tag{16}$$

The time series $\{AggArb_0, \ldots, AggArb_{T-1}\}$ is akin to the ETF one-period-lagged share growth series $\{g_0, \ldots, g_{T-1}\}$ and the time series $\{AggRet_1, \ldots, AggRet_T\}$ is akin to the ETF return series $\{r_1, \ldots, r_T\}$ in the share-growth-adjusted return calculation. We calculate the realized aggregate share-growth-adjusted return and compare it to the expected share-growth-adjusted return — the realized return falls 33 bps below the expected return and the realized return is statistically different than the expected return with a p-value of 0.1%. To put the effective fee (i.e., underperformance) of 33 bps into perspective, the mature ETF sample represents approximately 2.3 trillion dollars in assets at the end of 2016 and a 33 bps effective fee amounts to 7.7 billion dollars in underperformance in 2016 alone. Earlier results suggest that the correlation between arbitrage activity and subsequent returns has weakened in recent years. Repeating the analysis over two sub-samples, January 2007 - December 2011 and January 2012 - December 2016, the respective effective fees (p-values) are 55 bps (0.6%) and 7 bps (4.0%). The sub-sample analysis again demonstrates a weakening relation between ETF arbitrage activity in later periods, but one that is still statistically significant. Furthermore, a 7 bps effective fee still amounts to over 1.6 billion dollars in underperformance in 2016.

5 Conclusion

Weak-form market efficiency implies that all publicly observable information is impounded into prices. As a result, if markets are weak-form efficient, conditioning on observable arbitrage activity should not lead to predictable returns. Moreover, any predictability should be relatively short-lived. We find that weak-form efficiency does not hold for ETFs or their underlying assets, even at a monthly level. Our results show that when a large amount of money flows into ETFs, it leads to distortions in the ETFs and underlying assets prices. Thus, ETF investors tend to systematically buy (and sell) assets at the wrong prices.

Our results make several contributions. First, we show that *relative* price efficiency does not necessarily imply *absolute* price efficiency. Second, consistent with theoretical predictions and short-horizon results (Ben-David et al., 2017a), we find that ETFs transmit non-fundamental demand shocks to underlying assets. Importantly, we show that market participants do not quickly unwind these non-fundamental shocks, leading to monthly return predictability in both the cross-section and time-series of ETFs. Finally, we provide new evidence that active managers may be able to earn excess returns, as a group, because passive indexers systematically underperform. As such, our results provide a new counterpoint to Sharpe's well known arithmetic of active management. Overall, our results show that non-fundamental traders and arbitrageurs exert powerful impacts on the dynamics of asset prices. Surprisingly, despite arbitrageurs' trades being easily observable, market participants allow their price impacts to persist.

References

Agarwal, V., Hanouna, P., Moussawi, R., & Stahel, C. (2017). Do ETFs Increase the Commonality in Liquidity of Underlying Stocks? *Working Paper*.

Arif, S., Ben-Rephael, A., & Lee, C. (2016). What Drives Short-Sale Strategies? New Evidence from Mutual Fund Trades. *Working Paper*.

Baltussen, G., van Bekkum, S., & Da, Z. (2016). Indexing and Stock Market Serial Dependence Around the World. *Working Paper*.

Ben-David, I., Franzoni, F., & Moussawi, R. (2017a). Do ETFs Increase Volatility? *Journal of Finance*, *Forthcoming*.

Ben-David, I., Franzoni, F., & Moussawi, R. (2017b). Exchange Traded Funds (ETFs). *Annual Review of Financial Economics*, *9*, 165–185.

Bessembinder, H. (2015). Predictable ETF Order Flows and Market Quality. *Working Paper*.

Bhattacharya, A., & O'Hara, M. (2016). Can ETFs Increase Market Fragility? Effect of Information Linkages in ETF Markets. *Working Paper*.

Bhattacharya, U., Loos, B., Meyer, S., & Hackethal, A. (2016). Abusing ETFs. *Review of Finance*, *21*(3), 1217–1250.

BlackRock. (2014). ETP Landscape, Industry Highlights. *BlackRock Report*.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, *52*(1), 57-82.

Cella, C., Ellul, A., & Giannetti, M. (2013). Investors' Horizons and the Amplification of Market Shocks. *Review of Financial Studies*, hht023.

Cheng, M., & Madhavan, A. (2009). The Dynamics of Leveraged and Inverse Exchange-Traded Funds. *Journal of Investment Management*, *7*(4).

Coles, J., Heath, D., & Ringgenberg, M. C. (2017). On Index Investing. *Working Paper*.

Coval, J., & Stafford, E. (2007). Asset Fire Sales (and Purchases) in Equity Markets. *Journal of Financial Economics*, *86*(2), 479–512.

Da, Z., & Shive, S. (2016). Exchange-Traded Funds and Equity Return Correlations. *Work-*

ing Paper.

Davies, S. (2017). Speculation Sentiment. *Working Paper*.

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

Evans, R. B., Moussawi, R., Pagano, M. S., & Sedunov, J. (2016). ETF Failures-to-Deliver: Naked Short-Selling or Operational Shorting? *Working Paper*.

Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, *33*, 3-56.

Fama, E. F., & French, K. R. (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, *116*(1), 1-22.

Fulkerson, J. A., & Jordan, B. D. (2013). Reading Tomorrow's Newspaper: Predictability in ETF Returns. *Journal of Index Investing*, *4*, 23–31.

Glosten, L. R., Nallareddy, S., & Zou, Y. (2016). ETF Trading and Informational Efficiency of Underlying Securities. *Working Paper*.

Hombert, J., & Thesmar, D. (2014). Overcoming Limits of Arbitrage: Theory and Evidence. *Journal of Financial Economics*, *111*(1), 26–44.

Hou, K., Xue, C., & Zhang, L. (2016). A Comparison of New Factor Models. *Working Paper*.

Huang, S., Ringgenberg, M. C., & Zhang, Z. (2017). The Information in Fire Sales. *Working Paper*.

Israeli, D., Lee, C., & Sridharan, S. A. (2015). Is There a Dark Side to Exchange Traded Funds (ETFs)? An Information Perspective. *An Information Perspective (July 1, 2015)*.

Krause, T., Ehsani, S., & Lien, D. (2013). Exchange Traded Funds, Liquidity, and Market Volatility. *Working Paper*.

Lamont, O. A., & Thaler, R. H. (2003). Anomalies: The Law of One Price in Financial Markets. *The Journal of Economic Perspectives*, *17*(4), 191–202.

Lou, D. (2012). A Flow-Based Explanation for Return Predictability. *Review of Financial*

Studies, *25*(12), 3457–3489.

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

Madhavan, A. (2016b). Exchange-Traded Funds: An Overview of Institutions, Trading, and Impacts. *Annual Review of Financial Economics*, *6*, 311-341.

Malamud, S. (2015). A Dynamic Equilibrium Model of ETFs. *Working Paper*.

Pedersen, L. H. (2016). Sharpening the Arithmetic of Active Management. *Working Paper*.

Petajisto, A. (2011). The Index Premium and Its Hidden Cost for Index Funds. *Journal of Empirical Finance*, *18*, 271-288.

Petajisto, A. (2017). Inefficiencies in the Pricing of Exchange-Traded Funds. *Financial Analysts Journal*, *73*(1), 24–54.

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical association*, *89*(428), 1303–1313.

Pontiff, J. (1996). Costly Arbitrage: Evidence from Closed-End Funds. *The Quarterly Journal of Economics*, 1135–1151.

Sharpe, W. F. (1991). The Arithmetic of Active Management. *Financial Analysts Journal*, *47*(1), 7–9.

Shleifer, A. (1986). Do Demand Curves for Stocks Slope Down? *Journal of Finance*, *41*, 579-590.

Shleifer, A., & Summers, L. H. (1990). The Noise Trader Approach to Finance. *The Journal of Economic Perspectives*, *4*(2), 19–33.

Shleifer, A., & Vishny, R. W. (1997). The Limits of Arbitrage. *The Journal of Finance*, *52*(1), 35–55.

Staer, A. (2016). Fund Flows and Underlying Returns: The Case of ETFs. *Available at SSRN 2158468*.

Staer, A., & Sottile, P. (2016). Equivalent Volume and Comovement. *Working Paper*.

Figure 1: Price Dislocations and ETF Arbitrage Activity.

At $t = 0$, a small mispricing exists between the ETF share price (ETF_t) and the ETF NAV (NAV_t). At $t = 1$, an imbalanced demand shock generates a larger mispricing by pushing the ETF share price and the ETF NAV away from their initial values, with a larger impact on the ETF share price. At $t = 2$, arbitrageurs restore relative price efficiency, putting upward price pressure on the ETF NAV and downward price pressure on the ETF share price.

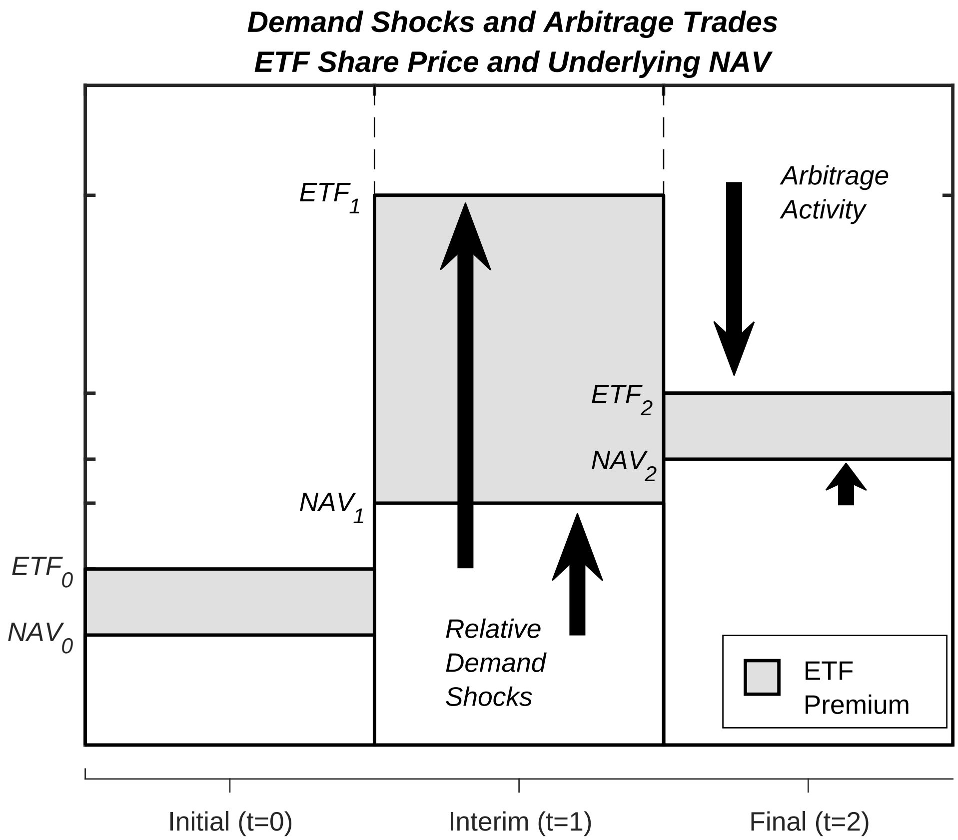

Figure 2: Cumulative Abnormal ETF Returns Around Top-Quintile Share Creation Events.

Raw ETF returns are calculated for top and bottom decile ETFs, based on monthly sorts, using the sample of only mature ETFs.

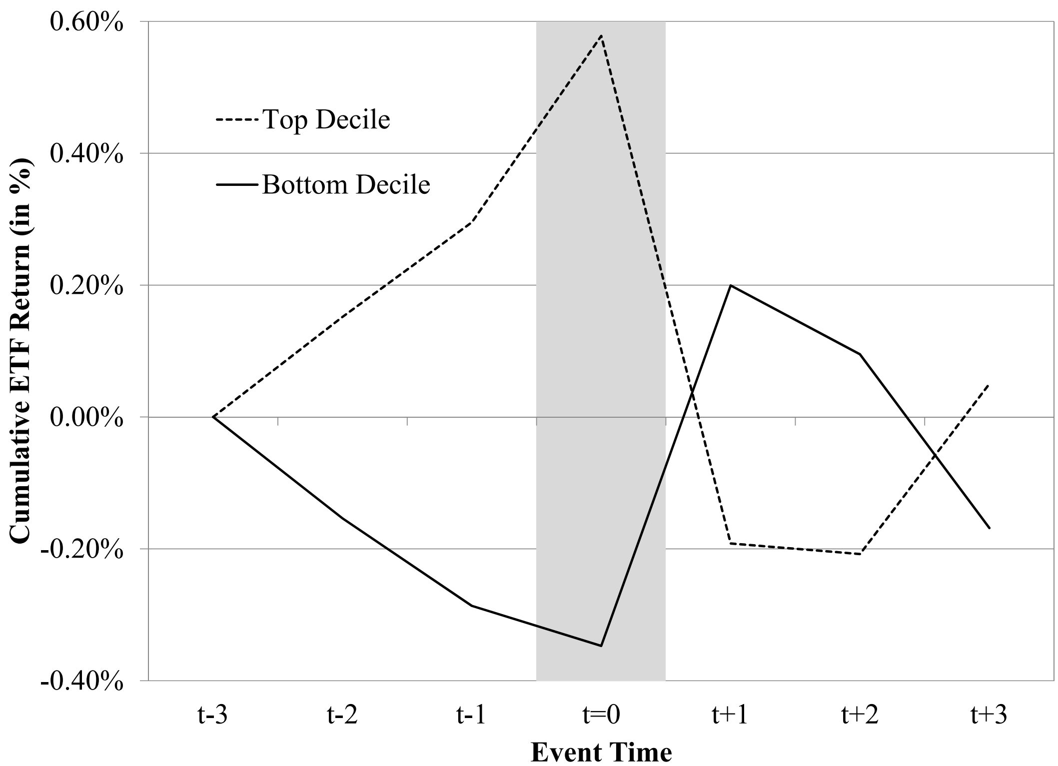

Table 1: ETFs Sample Per Year.

ETFs are included in the $50M+ sample from the first month in which end-of-month market capitilization exceeds $50 million. ETFs are considered mature from the first month in which creation or redemption activity was reported on at least 50% of trading days.

	All ETFs		$50M+ ETFs		Mature ETFs	
Year	Number	Market Cap	Number	Market Cap	Number	Market Cap
2007	581	$605	370	$603	124	$516
2008	682	$532	447	$530	178	$468
2009	772	$774	525	$770	227	$687
2010	927	$993	635	$988	270	$891
2011	1,131	$1,044	732	$1,039	331	$956
2012	1,208	$1,341	807	$1,336	360	$1,223
2013	1,299	$1,682	911	$1,677	408	$1,529
2014	1,412	$1,976	1,029	$1,971	439	$1,785
2015	1,589	$2,108	1,113	$2,103	516	$1,914
2016	1,707	$2,532	1,178	$2,526	564	$2,309

Table 2: ETF Summary Characteristics.

ETFs are included in the \$50M+ sample from the first month in which end-of-month market capitilization exceeds \$50 million. ETFs are considered mature from the first month in which creation or redemption activity was reported on at least 50% of trading days.

	\$50M+ ETFs	Mature ETFs
Average ETF Characteristics		
Shares Outstanding (millions)	29.8	59.6
Average Monthly Volume (millions)	35	76
Average Monthly Volume (percentage of shares out)	94.1%	157.0%
ETF Market Capitalization (billions)	\$1.7	\$3.5
Bid-Ask Spread	0.19%	0.10%
Short Interest Percentage	6.7%	11.5%
Percent of Active Days	21.7%	36.9%
Monthly Observations	88,324	38,648
Lipper Category Percentages		
General Equities	33.7%	33.3%
Sector-Based Equities	23.7%	28.5%
Bonds	17.8%	14%
Commodities	6.1%	6.9%
International	18.7%	17.2%

Table 3: Equity portfolio sorts based on ETF arbitrage activity.

The table displays one-month ahead excess returns and one-month ahead 4-factor portfolio alphas (calculated using the Fama-French 3 factors plus the momentum factor). Each month, we allocate ETF inflows and outflows to individual stocks (based on their weight in the ETF) to develop a measure of price pressure from ETF creation and redemption activity. We then sort stocks into deciles based on this measure, where portfolio one contains stocks with the highest outflows and portfolio ten contains stocks with the highest inflows. We also independently sort stocks into terciles based on their market capitalization and their bid-ask spread each period. Panel A displays results from sorts on ETF Arbitrage Activity for all stocks, Panel B displays results from dual independent sorts on ETF Arbitrage Activity and Market Capitalization, and Panel C displays results from dual independent sorts on ETF Arbitrage Activity and Bid-Ask Spread. The Long-Short column displays results from a long-short strategy that buys stocks in portfolio 1 and short sells stocks in portfolio 10. t-statistics calculated using standard errors clustered by year-month are shown below the estimates in italics. ***,**,* indicates statistical significance at the 1%, 5%, and 10% levels, respectively.

	One Month Ahead Excess Return			One Month Ahead 4-factor Alpha		
	Decile 1	Decile 10	Long-Short	Decile 1	Decile 10	Long-Short
Panel A: Sorts on ETF Arbitrage						
All Stocks	1.355***	0.744	0.611**	0.662***	0.059	0.602**
	(2.63)	(1.51)	(2.50)	(3.24)	(0.40)	(2.20)
Panel B: Dual Sorts on ETF Arbitrage and Bid-Ask Spread						
Low Bid-Ask Stocks	0.876*	0.537	0.34	0.222	-0.061	0.282
	(1.75)	(1.19)	(1.27)	(1.20)	(-0.39)	(1.02)
Medium Bid-Ask Stocks	1.253**	0.906	0.347	0.518***	0.128	0.391
	(2.24)	(1.62)	(1.45)	(2.78)	(0.72)	(1.48)
High Bid-Ask Stocks	1.504***	1.241**	0.264	0.764***	0.513**	0.252
	(2.79)	(2.28)	(0.85)	(2.77)	(2.47)	(0.69)
Panel C: Dual Sorts on ETF Arbitrage and Volume						
Low Volume Stocks	1.389***	0.756*	0.633**	0.779***	0.116	0.663**
	(3.07)	(1.65)	(2.33)	(3.94)	(0.63)	(2.39)
Medium Volume Stocks	1.437***	0.912**	0.525**	0.741***	0.304*	0.437
	(2.88)	(2.00)	(1.99)	(3.72)	(1.71)	(1.56)
High Volume Stocks	0.971	0.673	0.299	0.170	-0.099	0.269
	(1.53)	(1.15)	(0.92)	(0.66)	(-0.41)	(0.74)

Table 4: Panel regressions of equity returns on the prior month's ETF arbitrage activity.

The table displays one-month ahead excess equity returns (in percent) regressed on ETF creation and redemption activity according to the model:

$$Ret_{i,t+1} = \beta_1(ETFArb_{i,t}) + FE_i + FE_{j\times y} + \epsilon_{i,t+1},$$

where $Ret_{i,t+1}$ is the one-month ahead excess stock return (in percent) from CRSP, ETF $Arb_{i,t}$ is the price pressure in stock i in month t due to ETF Arbitrage Activity in that stock to meet creation and redemption demands, FE_i is firm fixed effects, and $FE_{j\times y}$ is industry $\times$ year fixed effects calculated using 1-digit SIC codes. Each month, we allocate ETF inflows and outflows to individual stocks (based on their weight in the ETF) to develop a measure of equity price pressure from ETF arbitrage activity associated with creations and redemptions. Columns (1) through (4) contain results from a continuous measure of ETF Arbitrage, while columns (5) through (8) use decile ranks. Each month, we sort stocks into ten portfolios based on ETF arbitrage activity, where portfolio one contains stocks with the highest outflows and portfolio ten contains stocks with the highest inflows. Models (3), (4), (7), and (8) examine sub-periods, as indicated at the bottom of the table. Firm and/or industry-year fixed effects are indicated at the bottom of the table. t-statistics calculated using standard errors clustered by firm and year-month are shown below the estimates in italics. ***,**,* indicates statistical significance at the 1%, 5%, and 10% levels, respectively.

Explanatory	Dependent Variable: One Month Ahead Excess Return (in %)							
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ETF Arb.	-0.015**	-0.013***	-0.015*	-0.012***				
	(-2.44)	(-2.98)	(-1.81)	(-3.31)				
Decile 2					-0.095	-0.112	-0.273	0.004
					(-0.53)	(-0.67)	(-1.21)	(0.01)
Decile 3					-0.355*	-0.374*	-0.487*	-0.311
					(-1.69)	(-1.85)	(-1.74)	(-1.06)
Decile 4					-0.446**	-0.455**	-0.617**	-0.359
					(-2.20)	(-2.31)	(-2.28)	(-1.24)
Decile 5					-0.519**	-0.536**	-0.661**	-0.503
					(-2.36)	(-2.56)	(-2.42)	(-1.61)
Decile 6					-0.606***	-0.630***	-0.787**	-0.547*
					(-2.64)	(-2.87)	(-2.48)	(-1.78)
Decile 7					-0.596**	-0.617***	-0.657**	-0.618*
					(-2.54)	(-2.70)	(-2.04)	(-1.90)
Decile 8					-0.674***	-0.688***	-0.663*	-0.733**
					(-2.84)	(-2.95)	(-1.96)	(-2.30)
Decile 9					-0.580**	-0.586**	-0.535	-0.638*
					(-2.40)	(-2.50)	(-1.53)	(-1.99)
Decile 10					-0.664***	-0.697***	-0.558	-0.798**
					(-2.70)	(-2.79)	(-1.56)	(-2.23)
Sample	2007-16	2007-16	2007-11	2012-16	2007-16	2007-16	2007-11	2012-16
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry-Year FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Adj. R^2	1.7%	4.3%	5.3%	4.0%	1.6%	4.3%	5.3%	3.9%
Observations	322,042	321,985	163,769	158,160	322,042	321,985	163,769	158,160

Table 5: Univariate portfolio sorts based on the prior month's ETF arbitrage activity.

Sorts are done within months. Panel A presents raw returns and Panel B presents four-factor portfolio alphas (calculated using the Fama-French three factors plus the momentum factor) . The first set of five columns compares the top and bottom deciles, whereas the second set compares the top decile against all other deciles averaged together.

Panel A: Baseline Results

	Raw Portfolio Alphas				Four-Factor Portfolio Alphas			
	Decile 1	Decile 10	Difference	t-statistic	Decile 1	Decile 10	Difference	t-statistic
$50M+ ETFs								
Equal-Weighted	0.442	-0.554	**0.996*** **	3.223	0.074	-0.986	**1.059*** **	3.373
Value-Weighted	0.558	-0.243	**0.831** **	2.042	0.170	-0.685	**0.855** **	2.106
Mature ETFs								
Equal-Weighted	0.681	-1.312	**1.993*** **	4.010	0.369	-1.774	**2.142*** **	4.275
Value-Weighted	0.712	-0.485	**1.196** **	2.203	0.339	-0.989	**1.328** **	2.450

Panel B: Sorted by Creation/Redemption Activity

	Raw Portfolio Alphas				Four-Factor Portfolio Alphas			
	Decile 1	Decile 10	Difference	t-statistic	Decile 1	Decile 10	Difference	t-statistic
Low Creation/Redemption Activity								
Equal-Weighted	0.431	0.048	**0.383**	1.117	0.038	-0.339	**0.377**	0.899
Value-Weighted	0.618	0.521	**0.097**	0.261	0.119	0.083	**0.036**	0.095
Medium Creation/Redemption Activity								
Equal-Weighted	0.750	-1.171	**1.921*** **	3.652	0.429	-1.471	**1.900*** **	3.545
Value-Weighted	0.860	-0.644	**1.504*** **	2.715	0.543	-1.039	**1.583*** **	2.822
High Creation/Redemption Activity								
Equal-Weighted	0.908	-1.743	**2.652*** **	3.003	0.608	-2.305	**2.913*** **	3.269
Value-Weighted	1.040	-0.785	**1.825** **	2.328	0.708	-1.328	**2.037*** **	2.632

Table 6: Panel regressions of monthly ETF returns on prior months' ETF arbitrage activity

The table displays panel regressions of monthly ETF returns (in percent) on measures of past creation and redemption activity according to the model:

$$Ret_{j,t+1} = \alpha + \sum_{d=1:10} \beta_d Decile_{j,d,t} + \Gamma X_t + \delta V_{j,t} + \alpha_j + \epsilon_{j,t+1},$$

where $Ret_{j,t+1}$ is the monthly return on ETF j including distributions, $Decile_{j,d,t}$ is an indicator variable for whether ETF j is in decile portfolio d in period t, X_t are factor returns, $V_{j,t}$ are ETF characteristics, and α_j are ETF fixed effects. The first five columns use deciles based on monthly-sorts of past share change as a percentage of shares outstanding. The last three columns use a standardized measure of past share change as a percentage of shares outstanding. The use of control variables, including ETF fixed effects, four-factor returns and lagged ETF characteristics (return, premium, market cap and volume) are indicated at the bottom of the table. t-statistcs calculated using standard errors clustered by firm and year-month are shown below the estimates in parentheses. ***,**,* indicates statistical significance at the 1%, 5%, and 10% levels, respectively.

Explanatory	Dependent Variable: One Month Ahead Return (in %)							
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Decile2	-0.47*	-0.52*	-0.52**				-0.46	-0.58**
	(-1.78)	(-1.94)	(-2.08)				(-0.86)	(-2.06)
Decile3	-0.47*	-0.54*	-0.54**				-0.47	-0.63**
	(-1.71)	(-1.94)	(-2.03)				(-0.81)	(-2.24)
Decile4	-0.60**	-0.61**	-0.59**				-0.72	-0.60**
	(-2.17)	(-2.20)	(-2.25)				(-1.31)	(-2.05)
Decile5	-0.26	-0.43	-0.41				-0.59	-0.40
	(-0.77)	(-1.40)	(-1.38)				(-1.07)	(-1.15)
Decile6	-0.46	-0.47	-0.43				-0.66	-0.38
	(-1.41)	(-1.48)	(-1.41)				(-1.05)	(-1.16)
Decile7	-0.42	-0.45	-0.41				-0.50	-0.41
	(-1.52)	(-1.60)	(-1.50)				(-0.85)	(-1.45)
Decile8	-0.52*	-0.55*	-0.51*				-0.54	-0.58*
	(-1.78)	(-1.88)	(-1.77)				(-0.89)	(-1.87)
Decile9	-0.65**	-0.66**	-0.62*				-0.79	-0.64*
	(-2.00)	(-2.02)	(-1.97)				(-1.37)	(-1.73)
Decile10	-1.58***	-1.56***	-1.54***				-1.95**	-1.44***
	(-3.32)	(-3.28)	(-3.47)				(-2.05)	(-2.99)
LagShareChange				-0.12	-0.10	-0.10		
				(-1.23)	(-1.24)	(-1.23)		
Sample	2007-16	2007-16	2007-16	2007-16	2007-16	2007-16	2007-11	2012-16
ETF Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Factor Returns	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Lagged Controls	No	No	Yes	No	No	Yes	Yes	Yes
Adjusted R^2	0.012	0.188	0.189	0.010	0.186	0.187	0.300	0.115
Observations	38,648	38,648	38,648	38,648	38,648	38,648	12,268	26,378

Table 7: Impact of ETF characteristics on relation between monthly ETF returns and ETF arbitrage activity

The table displays panel regressions of monthly ETF returns (in percent) on measures of past creation and redemption activity interacted with ETF characteristics according to the model:

$$\begin{aligned} Ret_{j,t+1} &= \alpha + \beta_1 Decile1_{j,t} + \beta_2 Decile1_{j,t} \times ETFCharacteristic_j + \beta_3 Decile10_{j,t} \\ &\quad + \beta_4 Decile10_{j,t} \times ETFCharacteristic_j + \Gamma X_t + \alpha_j + \epsilon_{j,t+1}, \end{aligned}$$

where $Ret_{j,t+1}$ is the monthly return on ETF j including distributions, $DecileD_{j,t}$ is an indicator variable for whether ETF j is in decile portfolio d in period t, X_t are factor returns, $ETFCharacteristic$ are ETF characteristics, and α_j are ETF fixed effects. The use of control variables, including ETF fixed effects and four-factor returns are indicated at the bottom of the table. t-statistcs calculated using standard errors clustered by firm and year-month are shown below the estimates in parentheses. ***,**,* indicates statistical significance at the 1%, 5%, and 10% levels, respectively.

Explanatory	Dependent Variable: One Month Ahead Return (in %)						
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Decile1	0.54**	-0.02	0.17	0.76*	0.62**	0.50*	0.61*
	(1.98)	(-0.13)	(0.89)	(1.94)	(2.06)	(1.82)	(1.95)
× Levered		1.55*					
		(1.89)					
× BroadEquity			1.02				
			(1.52)				
× SectorEquity				-0.68			
				(-1.54)			
× Bond					-0.90**		
					(-2.08)		
× Commodity						0.70	
						(0.63)	
× International							-0.51
							(-1.15)
Decile10	-1.02***	-0.21	-0.18	-1.31***	-1.15***	-1.13***	-1.14***
	(-3.30)	(-1.54)	(-0.96)	(-3.11)	(-3.32)	(-3.70)	(-3.25)
× Levered		-2.59***					
		(-2.93)					
× BroadEquity			-2.30***				
			(-3.32)				
× SectorEquity				1.02**			
				(2.27)			
× Bond					1.09**		
					(2.53)		
× Commodity						1.40	
						(1.33)	
× International							0.89*
							(1.87)
ETF Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Factor Returns	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Controls	No	No	No	No	No	No	No
Adjusted R^2	0.188	0.191	0.190	0.188	0.188	0.188	0.188
Observations	38,648	38,648	38,648	38,648	38,648	38,648	38,648

Table 8: Realized share-growth-adjusted returns **r** compared to simulated share-growth-adjusted returns.

Each panel reports the percentage of ETFs for which the realized share-growth-adjusted returns **r** are statistically different than the simulated share-growth-adjusted returns based on thresholds of 1% (99%), 2.5% (97.5%), 5% (95%), and 10% (90%). The first two columns report the results based on equal weights; the first column reports the fraction of realized share-growth-adjusted returns that are smaller than simulated share-growth-adjusted returns and the second column reports the fraction that are larger. The last two columns report the results based on average market capitalization share weights; the third column reports the value-weighted fraction of realized share-growth-adjusted returns that are smaller than simulated share-growth-adjusted returns and the fourth column reports the fraction that are larger. Panel A reports the results for the entire sample period of January 2007 - December 2016. Panel B presents the results over the first half of the sample period from January 2007 - December 2011 and Pancel C presents the results over the second half of the sample period from January 2012 - December 2016.

Panel A: Entire Sample				
Thresholds	Equal Weighted		Market Cap Share Weighted	
	$\% < p$	$\% > 1-p$	$\% < p$	$\% > 1-p$
1.00 %	3.16 %	0.49 %	14.58 %	0.30 %
2.50 %	6.31 %	3.16 %	17.42 %	4.35 %
5.00 %	11.65 %	6.80 %	26.18 %	5.34 %
10.00 %	17.72 %	12.38 %	31.26 %	9.62 %
$N = 412$				
Panel B: January 2007 - December 2011				
Thresholds	Equal Weighted		Market Cap Share Weighted	
	$\% < p$	$\% > 1-p$	$\% < p$	$\% > 1-p$
1.00 %	3.83 %	0.55 %	15.87 %	0.08 %
2.50 %	7.65 %	2.73 %	20.51 %	2.02 %
5.00 %	11.48 %	6.01 %	30.35 %	2.98 %
10.00 %	21.86 %	10.93 %	36.91 %	5.82 %
$N = 183$				
Panel C: January 2012 - December 2016				
Thresholds	Equal Weighted		Market Cap Share Weighted	
	$\% < p$	$\% > 1-p$	$\% < p$	$\% > 1-p$
1.00 %	0.99 %	0.49 %	0.21 %	0.05 %
2.50 %	2.96 %	0.74 %	3.52 %	0.14 %
5.00 %	7.14 %	2.22 %	16.56 %	1.69 %
10.00 %	13.05 %	6.40 %	19.48 %	3.31 %
$N = 406$				

Data Appendix for "ETF Arbitrage and Return Predictability"

DAVID C. BROWN, SHAUN WILLIAM DAVIES, AND MATTHEW C. RINGGENBERG[23]

This appendix provides details on the data used in our empirical analyses.

[23]Citation format: Brown, David C., Shaun William Davies, and Matthew C. Ringgenberg, Data Appendix for "ETF Arbitrage and Return Predictability," 2017, Working Paper.

A.1 ETF Sample

Our ETF universe combines the ETFs listed in CRSP with the list of ETFs pulled from http://www.etf.com/etfanalytics/etf-finder in April 2017. For CRSP, we identify ETFs by selecting securities with share code of 73. However, using CRSP data to identify ETFs results in 15 ETFs being double counted. Most of these instances are due to a change in ownership of the fund, i.e. the fund sponsor changes. These observations should be collapsed into a single time-series with a common PERMNO. In CRSP, the acquiring PERMNO data is populated in each of these cases. Table 9 details the 15 ETFs with their ticker, original PERMNO, new PERMNO and the transition date. In our analysis, we use data from the newer PERMNO starting on the transition date.

Table 9: ETFs with two PERMNOs in CRSP.

Ticker	Original PERMNO	New PERMNO	Transition Date
BBH	87433	13126	December 21, 2011
ERUS	12375	15160	January 26, 2015
EWU	83216	14907	September 29, 2014
FCG	92059	16075	May 2, 2016
OIH	88896	13129	December 21, 2011
PPH	89479	13130	December 21, 2011
QQQ	86755	91953	April 12, 2007
RKH	88311	13125	December 21, 2011
RTH	88993	13131	December 21, 2011
SMH	88236	13132	December 21, 2011
SPXH	13974	16153	July 18, 2016
TAN	92617	13208	February 15, 2012
TRSK	13975	16152	July 18, 2016
YMLI	13770	15928	February 22, 2016
YMLP	13295	15927	February 22, 2016

One additional ETF, EWRI, is populated with an acquiring PERMNO. In this case, EWRI was liquidated on January 26, 2016, and sharesholders were compensated with shares of the ETF RSP. If the two time-series were combined, RSP's assets would be overstated prior to January 26, 2016. As a result, the two ETFs should be treated as a separate ETFs, and no adjustment to the data is necessary.

In comparing the list of ETFs from CRSP and ETF.com, we find a problem in CRSP's identification of ETFs. While almost all ETFs are correctly identified with share code 73, 26 ETFs are incorrectly coded with share code 74 (closed-end funds). These ETFs have creation/redemption windows that are regularly open, and therefore should not be classified as closed-end funds. Table 10 lists these 26 ETFs' tickers and PERMNOs.

Table 10: ETFs listed as closed-end funds in CRSP.

Ticker	PERMNO
BNO	93425
BOIL	13032
CANE	13000
CMDT	14070
CORN	93424
CROC	13513
EUFX	13444
FXCH	13017
FXSG	13767
GLTR	12326
KOLD	13031
SOYB	13002
TAGS	13310
UGA	92580
UNL	93120
USL	92509
USO	91208
WEAT	13001
WITE	12445
CPER	13102
DNO	93036
OUNZ	14631
UHN	92637
UNG	91947
USAG	13368
USCI	12066

We collect data on our full sample of ETFs from CRSP and Bloomberg. Several issues arise in matching the CRSP ETF data to the Bloomberg ETF data. First, in 22 cases, CRSP CUSIP does not match to the Bloomberg CUSIP. In each of these cases, we have verified that the time series for prices, volumes and shares outstanding are consistent such that the two data sources are identifying the same ETF, but with different CUSIP values. Table 11 details the 22 ETFs and their identifying information. Second, many ETFs are listed in

Bloomberg, but data is not available in CRSP. The missing ETFs are typically either new or hold foreign assets.

Table 11: ETFs with different CUSIPs between Bloomberg and CRSP.

Ticker	PERMNO	Bloomberg CUSIP	CRSP CUSIP	CRSP NCUSIP
BGZ	92817	25460E88	25459Y37	25459W15
BMLA	16056	44053G30	75623U50	75623U50
DAX	15001	44053G20	75623U20	75623U20
FCGL	11182	25490K56	25490K34	25459W22
FINZ	13517	74348A18	74348A51	74348A51
IDXJ	13297	57060U16	92189F65	92189F65
JDST	14191	25490K54	25490K14	25490K14
JUNE	16032	13206187	28622M20	28622M20
MES	92732	57060U77	92189F85	92189F85
NGE	13842	37954Y66	37950E42	37950E42
QYLD	14354	44053G10	75623U10	75623U10
RUSS	12727	25460E82	25490K78	25490K78
SDOW	93255	74348A17	74347X11	74347X11
SOXS	93283	25460E83	25490K77	25490K77
SPDN	16128	25460E86	25490K21	25490K21
SQQQ	93268	74348A16	74348A41	74348A41
SRTY	93257	74348A15	74348A33	74348A33
TPYP	15467	56167N72	61177620	61177620
UCO	92842	74347W24	74347W32	74347W32
USMR	16335	44053G40	75623U60	75623U60
UVXY	13030	74347W23	74347W25	74347W25
WDRW	15612	25460E80	25459Y12	25459Y12

A.2 Split Adjustments

To join the Bloomberg and CRSP data, we must first ensure that both data sets are reporting comparable information. While Bloomberg split-adjusts shares outstanding, volume and price, CRSP provides raw values. Accordingly, we use the CRSP data, along with split dates provided by Bloomberg, to create split-adjusted time series of shares outstanding, volumes and prices. We retain the raw data and split adjustment factor as well for comparison to other data sources that provide raw and not split-adjusted values.

We also identify twelve split-adjustment errors in the Bloomberg data, which are detailed in Table 12. The errors are identified by comparing the shares before and after a split, and examining abnormal values. For abnormal values, we use CRSP volume and shares outstanding data to confirm the data or verify errors. For the identified errors, we provide a required multiplier that should be applied to the Bloomberg data to give correct values.

Table 12: ETFs with Split-Adjustment Errors.

Ticker	PERMNO	First Date	Last Date	Required Multiplier
QQQ	91953	March 10, 1999	March 17, 2000	1/2
SLYV	88608	September 29, 2000	September 21, 2005	1/3
IGE	89188	October 26, 2001	June 08, 2005	1/2
DWAQ	89749	May 01, 2003	July 18, 2003	1/4
PWC	89748	May 01, 2003	July 18, 2003	1/4
ADRA	89560	May 30, 2003	July 07, 2006	1/3
ADRA	89560	January 03, 2006	July 07, 2006	1/3
ADRE	89562	January 03, 2006	July 07, 2006	1/4
ADRD	89561	January 03, 2006	July 07, 2006	1/3
ADRU	89563	January 03, 2006	July 07, 2006	1/3
UYG	91795	February 01, 2007	April 14, 2010	10
ITOT	89988	July 24, 2008	July 24, 2008	2

A.3 Data Cleaning Process

In order to avoid winsorizing our data, we take several steps to clean data on shares outstanding, volumes, prices, net asset values (NAVs) and returns. The following subsections detail the data cleaning processes. As Bloomberg is less ubiquitous in ETF research, we include a list of data fields in Table 13. The Bloomberg data is collected using the Bloomberg Excel add-in.

A.3.1 Shares Outstanding

Because CRSP data often changes infrequently (most commonly at a monthly frequency), Bloomberg is primary source for shares outstanding data. If Bloomberg does not have any shares outstanding data for a particular ETF, we use CRSP instead If Bloomberg data is missing some observations of shares outstanding, we use a two-stage process to fill in the data. In many cases, shares outstanding is reported as missing when an ETF does not trade

Table 13: Bloomberg fields utilized in analysis.

Bloomberg Field	Short Description	Field Type
COUNTRY	Bloomberg country code of the issuer.	Static
CREATE_REDEEM_PROCESS	Process by which an authorized participant can create or redeem shares of the Exchange Traded Fund (ETF) with the issuer. This can be done "in-kind" with baskets of the underlying securities, by using cash or through a hybrid of both.	Static
CREATION_CUTOFF_TIME	Daily deadline to accept new creations of the Exchange Traded Fund (ETF) defined by the ETF issuer.	Static
DERIVATIVES_BASED	Indicates whether this is an Exchange Traded Fund (ETF) that uses futures, forwards, swaps or options as a primary method of achieving exposure.	Static
EQY_DVD_HIST_SPLITS	Historical stock split information for a given security.	Time Series
EQY_INST_PCT_SH_OUT	Percentage of Shares Outstanding held by institutions. Institutions include 13Fs, US and International Mutual Funds, Schedule Ds (US Insurance Companies) and Institutional stake holdings that appear on the aggregate level. Based on holdings data collected by Bloomberg.	Time Series
EQY_SH_OUT	Total current number of shares outstanding.	Time Series
ETF_UNDL_INDEX_TICKER	The index whose results the fund tries to mirror.	Static
FUND_ASSET_CLASS_FOCUS	Broad asset sector the fund will invest in as stated in the prospectus. Bloomberg asset classes include: equity, fixed income, mixed allocation, money market, real estate, commodity, specialty, private equity and alternative investment.	Static
FUND_CREATION_UNIT_SIZE	The unit size aggregation in which an authorized participant can create or redeem ETF shares.	Static
FUND_EXPENSE_RATIO	The amount investors pay for expenses incurred in operating a mutual fund (after any waivers).	Static
FUND_FLOW	Provides the calculated net value of all creation/redemption activity on a fund's primary listing.	Time Series
FUND_INCEPT_DT	Inception Date. The start date of the fund. It usually occurs after the initial subscription period.	Static
FUND_LEVERAGE	Leveraged funds are those that seek to achieve a daily return that is a multiple of, or inverse of, the daily return of their underlying markets/securities. They include products whose underlying indices are already leveraged.	Static
FUND_LEVERAGE_AMOUNT	Indicates the leverage amount. Leverage percent is calculated using Total Debt/Total Managed Assets.	Static
FUND_LEVERAGE_TYPE	Leveraged funds are those that seek to achieve a daily return that is a multiple of, or inverse of, the daily return of their associated index. Possible values are Long and Short.	Static
FUND_NET_ASSET_VAL	Net Asset Value (NAV). Determined by subtracting the liabilities from the portfolio value of the fund's securities, and dividing that figure by the number of outstanding shares.	Time Series
FUND_TYP	This classification refers to a funds structure. UIT (Unit Investment Trust); ETF (Exchange Traded Fund); ETC (Exchange Traded Commodity); ETN (Exchange Traded Notes);	Static
ID_CUSIP	Security identification number for the U.S. and Canada.	Static
ID_CUSIP_8_CHR	Security identification number for the U.S. and Canada.	Static
INDEX_WEIGHTING_METHODOLOGY	Process the Exchange Traded Fund (ETF) uses to weight the holdings of each security in the underlying portfolio based on predetermined criteria.	Static
INVERSE_FUND_INDICATOR	Represents whether this is an Exchange Traded Fund (ETF) that profits from a decline in the value of its underlying assets.	Static
INVESTS_IN_PHYSICAL_COMMODITIES	Represents whether this is an Exchange Traded Fund (ETF) that invests in the physical commodity.	Static
PX_ASK	Last price for the security.	Time Series
PX_BID	Highest price an investor will accept to pay for a security.	Time Series
PX_LAST	Lowest price an investor will accept to sell a security.	Time Series
REBALANCING_FREQUENCY	Frequency at which the security is rebalanced.	Static
REPLICATION_STRATEGY	Represents whether this is an Exchange Traded Fund (ETF) that tracks its index. Possible values are Full, Optimized, Derivative, Blend.	Static
SECURITIES_LENDING	Indicates whether an exchange traded fund (ETF) currently engages in or is eligible to lend out securities. However, if an ETF issuer is eligible for securities lending, but it is informed that they do not engage in securities lending, this field will show 'No'.	Static
SECURITY_LENDING_PURPOSE	Provides information concerning the purpose of the security lending program of the exchange-traded fund (ETF). Possible returns are Revenue Generation, Regulatory Collateralization, Unknown and Not Applicable.	Static
TOT_RETURN_INDEX_GROSS_DVDS	One day total return index. Gross dividends are used.	Time Series
VOLUME	Total number of shares traded on a security on the current day.	Time Series

on a particular day. For these cases, it appears appropriate to fill in the data with the prior day's value. We therefore first replace missing values of shares outstanding with lagged values from up to 5 trading days before the missing data.

If shares outstanding data is still missing, but CRSP has data for those missing days, we replace the missing values with CRSP data with an adjustment. Often, CRSP and Bloomberg do not agree on shares outstanding both before and after the missing data (possibly due to stale reporting in CRSP). Accordingly, we average the ratio of shares outstanding before and after the missing data, and apply this ratio to the CRSP data when filling in the missing observations. For example, if CRSP reports 80% as many shares outstanding as Bloomberg before the missing data, and CRSP and Bloomberg match after the missing data, then the missing data is filled based on 90% of the CRSP data. If the filling process would introduce a change in shares outstanding of more than 5% of shares outstanding and greater than 100,000 shares, then the data is left as missing.

A.3.2 Volume

CRSP is our primary source for volume data.If CRSP does not have any volume data for a particular ETF, we use Bloomberg instead. If CRSP data is missing some observations of volume, we attempt to use Bloomberg to fill in the data. However, before replacing the missing observations with Bloomberg data, we compare how well CRSP and Bloomberg data match when volume data is available from both sources. For a particular ETF, as long as differences in volume are less than 5% of each other on average, then the missing CRSP data is replaced with Bloomberg data for all missing observations for that ETF. If the differences in volume are greater than 5%, then we fill in the missing data using the Bloomberg data with the same ratio-based adjustment as is used for shares outstanding. Finally, if CRSP reports zero shares and Bloomberg reports positive volume, we replace the CRSP value with the Bloomberg data.

A.3.3 Price

CRSP is our primary source for price data. If CRSP does not have any price data for a particular ETF, we use Bloomberg instead. If CRSP data is missing some observations of price, we attempt to use Bloomberg to fill in the data. However, before replacing the missing observations with Bloomberg data, we compare how well CRSP and Bloomberg data match when price data is available from both sources. For a particular ETF, as long as differences in price are less than 0.1% of each other on average, then the missing CRSP data is replaced with Bloomberg data for all missing observations for that ETF. However, if this filling process would introduce a price change of greater than 5%, then the missing data is not replaced. Finally, if missing price data remains, it is filled with lagged prices from up to 5 days before.

A.3.4 NAV

Bloomberg is our only source of NAV data. When values are missing, we fill in missing values with lagged values from up to 5 days before.

A.3.5 Returns

CRSP is our primary source for return data. If CRSP returns are missing, we fill in the missing observations based on the return implied by the price data (which has already been filled). We also examine potential errors in CRSP returns. If the absolute CRSP return is greater than 5% and the Bloomberg total return and price change are within 1%, then we replace the CRSP return with return implied by the price change.

A.3.6 Subsequent changes

We make several additional changes to the data set once each data series has been filled. First, if shares outstanding and price are both non-missing, but volume is missing, we set volume equal to zero. Second, we correct prices, shares outstanding and NAV that appear to be erroneous observations. If the one-day change in price, shares outstanding or NAV is less than -50% or more than 100%, and if the values on the two surrounding days are within 5% of one another, then we replace the suspect observation with the lagged value. Third, once all of the data filling is completed, we identify months that still have at least one day with missing data. We consider these months' observations to be unreliable and exclude them from our sample. However, we do not eliminate subsequent months provided that the missing data does not impact the data required for those subsequent months.

A.4 Data Assembly

We begin with our cleaned daily data set, create several additional measures, and supplement our data with several other data sources before reaching our final monthly sample.

A.4.1 Activity Measures

Our study focuses on the observed creation and redemption activity that signifies arbitrage activity surrounding ETFs and their underlying assets. Accordingly, we create several sub-samples of the data that exclude smaller and less-active ETFs. Our first sub-sample is based on ETF market capitalization. We include an ETF in our $50 million sample from the month-end in which it first had $50 million in market capitalization. As an example, BND reached $50 million in assets on April 26, 2007. We therefore include BND in our sample from the end of April and onward. Note that we do include ETFs in our sample that subsequently drop below $50 million in assets.[24] Our second sub-sample is based on the frequency of creations or redemptions within a month. Starting from our $50 million sample, we consider an ETF to be mature when at least 50% of the days within a month had either creation or redemption activity, i.e., the shares outstanding changed on at least 50% of days. For example, BND first had creation or redemption activity on at least 50% of days in January 2009. Because we used lagged values of changes in shares outstanding in our tests, we include ETFs in our mature sample from the month after they first become active. Thus, BND is included in our sample from February 2009 and onward.

A.4.2 Creating Monthly ETF Data

We join several additional data sources to our cleaned daily data set. First, we include static ETF data from Bloomberg on leverage and replication method. We also include static ETF data from the CRSP Mutual Fund Data on Lipper codes. We hand classify each Lipper code into one of the following five broad categories: Broad Equity, Sector Equity, Commodities, Bonds, International.[25] Second, we include daily 3-factor, 3-factor plus momentum and 5-factor returns from Ken French's website. Third, we include measures of bid-ask spread, short interest percent, the Amihud liquidity measure and retail trading volume, all at a daily frequency. Fourth, we include a random number for each daily observation. This random number can be added to the change in shares outstanding to serve as a tie-breaker for separating ETFs into test portfolios for asset pricing tests. While the added numbers are small enough to not affect the order of observations, and thus will not affect the extreme portfolios, it does allow separation of ETFs with zero shares outstanding change during a month.

[24] Eliminating ETF-months for which the end-of-month market capitalization is less than $50 million gives qualitatively similar results to those presented.

[25] For ETFs without CRSP data, we hand classify these funds based on internet searches.

We then aggregate our daily data set to monthly observations. For volume measures, we add up the daily observations within each month. For ETF and factor returns, we accumulate the returns to a total monthly return (including any distributions). We use month-end (last trading day) values to calculate share changes, premia of prices over net asset values (NAV), market capitalizations, and the associated changes in those variables. We also measures bid-ask spreads, short interest percent, and the Amihud liquidity measure based on month-end values.

As a final step, we include several filters to ensure our sample is consistent across tests. First, we required that each monthly observation in the sample has data for the change in shares outstanding, ETF and factor returns, volume, premium, market capitalization and the associated lagged values. Second, we required that no volume, price, shares outstanding or NAV data is missing for any days within the month. Finally, we require that each ETF has at least two observations in the sample.

A.4.3 Final Sample

Our final sample includes $2,196$ ETFs spanning from 2007 through 2016. Yearly sample details are provided in "ETF Arbitrage and Return Predictability." To aid in replication of our analyses and facilitate future research, we provide a complete list of these ETFs, the dates at which they reach $50 million in assets and are classified as "mature", and relevant identifiers for linking to Bloomberg, CRSP stock-level and CRSP mutual-fund-level data. The data will be available by the end of 2017 at www.davidclaytonbrown.com.

A.5 Data Sources and Download Dates

To aid in replication, Table 14 details the raw datasets we use in our analysis, as well as the dates the data was acquired.

Table 14: Bloomberg fields utilized in analysis.

Dataset	Source	Download Date	Short Description
Daily Data	Bloomberg	7/21/2017	Prices, volumes, shares outstanding, returns, net asset values
Split Dates	Bloomberg	7/21/2017	Dates of stock splits and adjustment ratios
ETF Characteristics	Bloomberg	7/21/2017	ETF leverage, ETF type, redemption details
Daily Stock Data	CRSP	7/18/2017	Prices, volumes, shares outstanding, returns
Mutual Fund Summary	CRSP	7/21/2017	Lipper codes, ETF/ETN identifier
Mutual Fund Portfolio Holdings	CRSP	7/25/2017	Shares held in individual stocks by ETF-date
Factor Returns	Ken French	6/10/2017	Daily 3-Factor, 5-Factor and Momentum Factor returns
ETF Identifiers	Created	9/6/2017	Links beteween CUSIP, ticker, PERMNO, FUNDNO and FundID (our unique identifier)
Broad Lipper Categories	Created	9/7/2017	Hand classified based on CRSP Lipper code data
Random Numbers	Created	9/1/2017	Small random numbers to ensure even distribution in deciles